UNIT 3 DATA PREPARATION FOR ANALYSIS

3.0 INTRODUCTION

In the previous unit of this Block, you were introduced to the basic concepts of conditional probability, Bayes Theorem and probability distribution including Binomial and Normal distributions. The Unit also introduces you to the concept of the sampling distribution, central limit theorem and statistical hypothesis testing. This Unit introduces you to the process of data preparation for Data Analysis. Data preparation is one of the most important processes, as it leads to good quality data, which will result in accurate results of the data analysis. This unit covers data selection, cleaning, curation, integration, and knowledge discovery from the stated data. In addition, this unit gives you an overview of data quality and how its preparation for analysis is done. You may refer to further readings for more details on these topics.

3.1 OBJECTIVES

After finishing this unit, you will be to be able to:

- Describe the meaning of "data quality"
- Explain basic techniques for data preprocessing
- Use the technique of data selection and extraction
- Define data curation and data integration
- Describe the knowledge discovery.

3.2 NEED FOR DATA PREPARATION

In the present time, data is one of the key resources for a business. Data is processed to create information; information is integrated to create knowledge. Since knowledge is power, it has evolved into a modern currency, which is valued and traded between parties. Everyone wants to discuss the knowledge and benefits they can gain from data. Data is one of the most significant resources available to marketers, agencies, publishers, media firms, and others today for a reason. But only high-quality data is useful. We can determine a data set's reliability and suitability for decision-making by looking at its quality. Degrees are frequently used to gauge this quality. The usefulness of the data for the intended purpose and its completeness, accuracy, timeliness, consistency, validity, and uniqueness are used to determine the data's quality. In simpler terms, data quality refers to how accurate and helpful the data are for the task at hand. Further, data quality also refers to the actions that apply the necessary quality management procedures and methodologies to make sure the data is useful and actionable for the data consumers. A wide range of elements, including accuracy, completeness, consistency, timeliness, uniqueness, and validity, influence data quality. Figure 1 shows the basic factors of data quality.

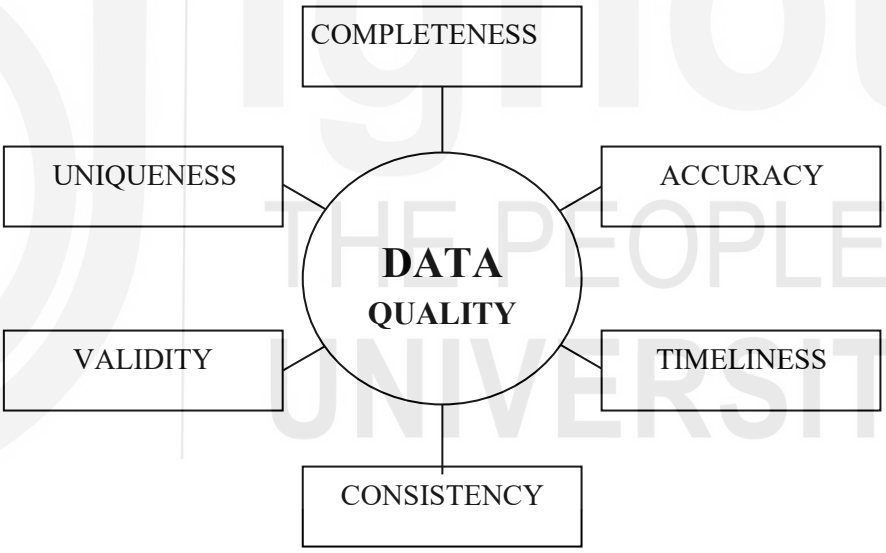

Figure 1: Factors of Data Quality

These factors are explained below:

- **Accuracy** - The data must be true and reflect events that actually take place in the real world. Accuracy measures determine how closely the figures agree with the verified right information sources.
- **Completeness** - The degree to which the data is complete determines how well it can provide the necessary values.
- **Consistency** - Data consistency is the homogeneity of the data across applications, networks, and when it comes from several sources. For example, identical datasets should not conflict if they are stored in different locations.

- **Timeliness** - Data that is timely is readily available whenever it is needed. The timeliness factor also entails keeping the data accurate; to make sure it is always available and accessible and updated in real-time.
- **Uniqueness** - Uniqueness is defined as the lack of duplicate or redundant data across all datasets. The collection should contain zero duplicate records.
- **Validity** - Data must be obtained in compliance with the firm's defined business policies and guidelines. The data should adhere to the appropriate, recognized formats, and all dataset values should be within the defined range.

Consider yourself a manager at a company, say XYZ Pvt Ltd, who has been tasked with researching the sales statistics for a specific organization, say ABC. You immediately get to work on this project by carefully going through the ABC company's database and data warehouse for the parameters or dimensions (such as the product, price, and units sold), which may be used in your study. However, your enthusiasm suffers a major problem when you see that several of the attributes for different tuples do not have any recorded values. You want to incorporate the information in your study on whether each item purchased was marked down, but you find that this data has not been recorded. According to users of this database system, the data recorded for some transactions were containing mistakes, such as strange numbers and anomalies.

The three characteristics of data quality—accuracy, completeness, and consistency—are highlighted in the paragraph above. Large databases and data warehouses used in the real world frequently contain inaccurate, incomplete, and inconsistent data. What may be the causes of such erroneous data in the databases? There may be problems with the data collection tools, which may result in mistakes during the data-entering process. Personal biases, for example, when users do not want to submit personal information, they may purposefully enter inaccurate data values for required fields (for example, by selecting the birthdate field's presented default value of "January 1"). Disguised missing data is what we call this. There may also be data transfer errors. The use of synchronized data transit and consumption may be constrained by technological limitations, such as a short buffer capacity. Unreliable data may result from differences in naming conventions, data codes, or input field types (e.g., date). In addition, cleansing of data may be needed to remove duplicate tuples.

Incomplete data can be caused by a variety of circumstances. Certain properties, such as customer information for sales transaction data, might not always be accessible. It is likely that some data was omitted since it was not thought to be important at the time of input. A misinterpretation or malfunctioning technology could prevent the recording of essential data. For example, the data, which did not match the previously stored data, was eliminated. Furthermore, it is likely that the data's past alterations or histories were not documented. In particular, for tuples with missing values for some properties, it could be required to infer missing data.

3.3 DATA PREPROCESSING

Preprocessing is the process of taking raw data and turning it into information that may be used. Data cleaning, data integration, data reduction and data transformation, and data discretization are the main phases of data preprocessing (see Figure 2).

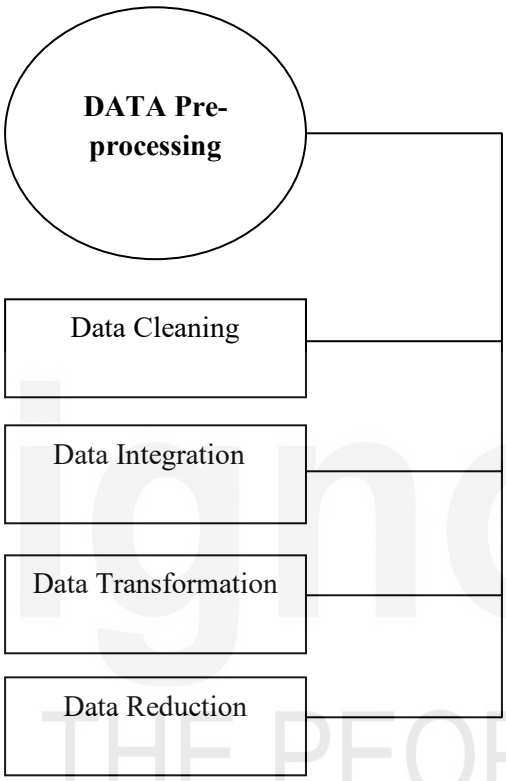

Figure 2: Data pre-processing

3.3.1 Data Cleaning

Data cleaning is an essential step in data pre-processing. It is also referred to as scrubbing. It is crucial for the construction of a good model. The step that is required but frequently overlooked by everyone is data cleaning. Real-world data typically exhibit incompleteness, noise, and inconsistency. In addition to addressing discrepancies, this task entails filling in missing numbers, smoothing out noisy data, and eliminating outliers. Errors are decreased, and data quality is enhanced via data cleansing. Although it might be a time-consuming and laborious operation, it is necessary to fix data inaccuracies and delete bad entries.

a. **Missing Values**

Consider you need to study customer and sales data for ABC Company. As you pointed out, numerous tuples lack recorded values for a number of characteristics, including customer income. The following techniques can be used to add the values that are lacking for this attribute.

i. **Ignore the tuple**: Typically, this is carried out in the absence of a class label (assuming the task involves

classification). This method is particularly detrimental when each attribute has a significantly different percentage of missing values. By disregarding the remaining characteristics in the tuple, we avoid using their values.

ii. **Manually enter the omitted value**: In general, this strategy is time-consuming and might not be practical for huge data sets with a substantial number of missing values.

iii. **Fill up the blank with a global constant**: A single constant, such as "Unknown" or "−∞", should be used to replace all missing attribute values. If missing data are replaced with, say, "Unknown," the analysis algorithm can mistakenly think that they collectively comprise valid data. So, despite being simple, this strategy is not perfect.

iv. **To fill in the missing value, use a measure of the attribute's central tendency (such as the mean or median)**: The median should be used for skewed data distributions, while the mean can be used for normal (symmetric) data distributions. Assume, for instance, that the ABC company's customer income data distribution is symmetric and that the mean income is INR 50,000/-. Use this value to fill in the income value that is missing.

v. **For all samples that belong to the same class as the specified tuple, use the mean or median**: For instance, if we were to categorize customers based on their credit risk, the mean income value of customers who belonged to the same credit risk category as the given tuple might be used to fill in the missing value. If the data distribution is skewed for the relevant class, it is best to utilize the median value.

vi. **Fill in the blank with the value that is most likely to be there**: This result can be reached using regression, inference-based techniques using a Bayesian formalization, or decision tree induction. As an example, using the other characteristics of your data's customers, you may create a decision tree to forecast the income's missing numbers.

b. **Noisy Data**

Noise is the variance or random error in a measured variable. It is possible to recognize outliers, which might be noise, employing tools for data visualization and basic statistical description techniques (such as scatter plots and boxplots). How can the data be "smoothed" out to reduce noise given a numeric property, like price, for example? The following are some of the data-smoothing strategies.

i. **Binning:** Binning techniques smooth-sorted data values by looking at their "neighbourhood" or nearby values. The values that have been sorted are divided into various "buckets" or bins. Binding techniques carry out local smoothing since they look at the values' surroundings. When smoothing by bin means, each value in the bin is changed to the bin's mean value. As an illustration, suppose a bin contains three numbers 4, 8 and 15. The average of these three numbers in the bin is 9.

Consequently, the value nine replaces each of the bin's original values.

Similarly, smoothing by bin medians, which substitutes the bin median for each bin value, can be used. Bin boundaries often referred to as minimum and maximum values in a specific bin can also be used in place of bin values. This type of smoothing is called smoothing by bin boundaries. In this method, the nearest boundary value is used to replace each bin value. In general, the smoothing effect increases with increasing breadth. As an alternative, bins may have identical widths with constant interval ranges of values.

ii. **Regression:** Regression is a method for adjusting the data values to a function and may also be used to smooth out the data. Finding the "best" line to fit two traits (or variables) is the goal of linear regression, which enables one attribute to predict the other. As an extension of linear regression, multiple linear regression involves more than two features and fits the data to a multidimensional surface.

iii. **Outlier analysis**: Clustering, for instance, the grouping of comparable values into "clusters," can be used to identify outliers. It makes sense to classify values that are outliers as being outside the set of clusters.

iv. **Data discretization**, a data transformation and data reduction technique, is an extensively used data smoothing technique. The number of distinct values for each property is decreased, for instance, using the binning approaches previously discussed. This functions as a form of data reduction for logic-based data analysis methods like decision trees, which repeatedly carry out value comparisons on sorted data. Concept hierarchies are a data discretization technique that can also be applied to smooth out the data. The quantity of data values that the analysis process must process is decreased by a concept hierarchy. For example, the price variable, which represents the price value of commodities, may be discretized into “lowly priced”, “moderately priced”, and “expensive” categories.

Steps of Data Cleaning

The following are the various steps of the data cleaning process.

1. **Remove duplicate or irrelevant observations-** Duplicate data may be produced when data sets from various sources are combined, scraped, or data is obtained from clients or other departments.

2. **Fix structural errors-**When measuring or transferring data; you may come across structural mistakes such as unusual naming practices, typographical errors, or wrong capitalization. Such inconsistencies may lead to mislabeled categories or classes. For instance, "N/A" and "Not Applicable", which might be present on any given document, may create two different classifications. Rather, they should be studied under the same heading or missing values.

3. **Managing Unwanted outliers** -Outliers might cause problems in certain models. Decision tree models, for instance, are more robust to outliers than linear regression models. In general, we should not eliminate outliers unless there is a compelling reason to do so. Sometimes removing them can improve performance, but not always. Therefore, the outlier must be eliminated for a

good cause, such as suspicious measurements that are unlikely to be present in the real data.

4. **Handling missing data**-Missing data is a deceptively difficult issue in machine learning. We cannot just ignore or remove the missing observation. They must be carefully treated since they can indicate a serious problem. Data gaps resemble puzzle pieces that are missing. Dropping it is equivalent to denying that the puzzle slot is there. It is like trying to put a piece from another puzzle into this one. Furthermore, we need to be aware of how we report missing data. Instead of just filling it in with the mean, you can effectively let the computer choose the appropriate constant to account for missingness by using this flagging and filling method.

5. **Validate and QA**-You should be able to respond to these inquiries as part of fundamental validation following the data cleansing process, for example:

- Does the data make sense?
- Does the data abide by the regulations that apply to its particular field?
- Does it support or refute your hypothesis? Does it offer any new information?
- Can you see patterns in the data that will support your analysis?
- Is there a problem with the data quality?

Methods of Data Cleaning

The following are some of the methods of data cleaning.

1. **Ignore the tuples:** This approach is not particularly practical because it can only be used when a tuple has multiple characteristics and missing values.
2. **Fill in the missing value:** This strategy is neither practical nor very effective. Additionally, the process could take a long time. The missing value must be entered into the approach. The most common method for doing this is by hand, but other options include attribute mean or using the value with the highest probability.
3. **Binning method:** This strategy is fairly easy to comprehend. The values nearby are used to smooth the sorted data. The information is subsequently split into a number of equal-sized parts. The various techniques are then used to finish the task.
4. **Regression:** With the use of the regression function, the data is smoothed out. Regression may be multivariate or linear. Multiple regressions have more independent variables than linear regressions, which only have one.
5. **Clustering:** The group is the primary target of this approach. Data are clustered together in a cluster. Then, with the aid of clustering, the outliers are found. After that, the comparable values are grouped into a "group" or "cluster".

3.3.2 Data Integration

Data from many sources, such as files, data cubes, databases (both relational and non-relational), etc., must be combined during this procedure. Both homogeneous and heterogeneous data sources are possible. Structured,

unstructured, or semi-structured data can be found in the sources. Redundancies and inconsistencies can be reduced and avoided with careful integration.

a. **Entity Identification Problem:** Data integration, which gathers data from several sources into coherent data storage, like data warehousing, will likely be required for your data analysis project. Several examples of these sources include various databases, data cubes, and flat files.
   During data integration, there are a number of things to consider. Integration of schemas and object matching might be challenging. How is it possible to match comparable real-world things across different data sources? The entity identification problem describes this. How can a computer or data analyst be sure that a client's ID in one database and their customer number in another database refer to the same attribute? Examples of metadata for each attribute include the name, definition, data type, permissible range of values, and null rules for handling empty, zero, or null values. Such metadata can be used to avoid errors during the integration of the schema. The data may also be transformed with the aid of metadata. For example, in two different instances of data of an organization, the code for pay data might be "H" for high income and "S" for small income in one instance of the database. The same pay code in another instance of a database maybe 1 and 2.
   When comparing attributes from one database to another during integration, the data structure must be properly considered. This is done to make sure that any referential constraints and functional dependencies on attributes present in the source system are also present in the target system. For instance, a discount might be applied to the entire order by one system, but only to certain items by another. If this is not found before integration, things in the target system can be incorrectly dismissed.
b. **Redundancy and Correlation Analysis:** Another crucial problem in data integration is redundancy. If an attribute (like annual income, for example) can be "derived" from another attribute or group of data, it may be redundant. Inconsistent attributes or dimension names can also bring redundancies in the final data set.

Correlation analysis can identify some redundancies. Based on the available data, such analysis can quantify the strength of relationships between two attributes. We employ the chi-square test ($\chi 2$) for finding relationships between nominal data. Numeric attributes can be analyzed using the correlation coefficient and covariance, which look at how one attribute's values differ from those of another.

c. **Tuple Duplication:** Duplication should be identified at the tuple level in addition to being caught between attributes (e.g., when, for a particular unique data entry case, there are two or more identical tuples). Additional data redundant sources include – the use of denormalized tables, which are frequently used to increase performance by reducing joins; faulty data entry or updating some (not all) redundant data occurrences, etc. Inconsistencies frequently appear between different duplicates. For example, there may be inconsistency as the same purchaser's name may appear with multiple addresses within the purchase order database. This might happen if a database for purchase orders has attributes for the buyer's name and address rather than a foreign key to this data.

d. **Data Value Conflict Detection and Resolution:** Data value conflicts must be found and resolved as part of data integration. As an illustration,
   attribute values from many sources may vary for the same real-world thing.

Variations in representation, scale, or encoding may be the cause of this. In one system, a weight attribute might be maintained in British imperial units, while in another, metric units. For a hotel chain, the cost of rooms in several cities could include various currencies, services (such as a complimentary breakfast) and taxes. Similarly, every university may have its own curriculum and grading system. When sharing information among them, one university might use the quarter system, provide three database systems courses, and grade students from A+ to F, while another would use the semester system, provide two database systems courses, and grade students from 1 to 10. Information interchange between two such universities is challenging because it is challenging to establish accurate course-to-grade transformation procedures between the two universities. An attribute in one system might be recorded at a lower abstraction level than the "identical" attribute in another since the abstraction level of attributes might also differ. As an illustration, an attribute with the same name in one database may relate to the total sales of one branch of a company, however, the same result in another database can refer to the company's overall regional shop sales.

3.3.3 Data Reduction

In this phase, data is trimmed. The number of records, attributes, or dimensions can be reduced. When reducing data, one should keep in mind that the outcomes from the reduced data should be identical to those from the original data. Consider that you have chosen some data for analysis from *ABC Company's* data warehouse. The data set will probably be enormous! Large-scale complex data analysis and mining can be time-consuming, rendering such a study impractical or unfeasible. Techniques for data reduction can be applied to create a condensed version of the data set that is considerably smaller while meticulously retaining the integrity of the original data. In other words, mining the smaller data set should yield more useful results while effectively yielding the same analytical outcomes. This section begins with an overview of data reduction tactics and then delves deeper into specific procedures. Data compression, dimensionality reduction, and numerosity reduction are all methods of data reduction.

a. **Dimensionality reduction** refers to the process of lowering the number of random variables or qualities. Principal components analysis and wavelet transformations are techniques used to reduce data dimensions by transforming or rescaling the original data. By identifying and eliminating duplicated, weakly relevant, or irrelevant features or dimensions, attribute subset selection is a technique for dimensionality reduction.
b. **Numerosity reduction** strategies substitute different, more compact forms of data representation for the original data volume. Both parametric and non-parametric approaches are available. In parametric techniques, a model is employed to estimate the data, which frequently necessitates the maintenance of only the data parameters rather than the actual data. (Outliers may also be stored.) Examples include log-linear models and regression. Nonparametric methods include the use of histograms, clustering, sampling, and data cube aggregation to store condensed versions of the data.
c. **Transformations** are used in **data compression** to create a condensed or "compressed" version of the original data. Lossless data compression is used when the original data can be recovered from the compressed data without

any information being lost. Alternatively, lossy data reduction is employed when we can only precisely retrieve a fraction of the original data. There are a number of lossless string compression algorithms; however, they typically permit only a small amount of data manipulation. Techniques for reducing numerosity and dimensions can also be categorized as data compression methods.

There are other additional structures for coordinating data reduction techniques. The time saved by analysis on a smaller data set should not be "erased" or outweighed by the computational effort required for data reduction.

Data Discretization: - It is regarded as a component of data reduction. The notional qualities take the place of the numerical ones. By converting values to interval or concept labels, data discretization alters numerical data. These techniques enable data analysis at various levels of granularity by automatically generating concept hierarchies for the data. Binding, histogram analysis, decision tree analysis, cluster analysis, and correlation analysis are examples of discretization techniques. Concept hierarchies for nominal data may be produced based on the definitions of the schema and the distinct attribute values for every attribute.

3.3.4 Data Transformation

This procedure is used to change the data into formats that are suited for the analytical process. Data transformation involves transforming or consolidating the data into analysis-ready formats. The following are some data transformation strategies:

a. **Smoothing**, which attempts to reduce data noise. Binning, regression, and grouping are some of the methods.
b. **Attribute construction** (or *feature construction*), wherein, in order to aid the analysis process, additional attributes are constructed and added from the set of attributes provided.
c. **Aggregation**, where data is subjected to aggregation or summary procedures to calculate monthly and yearly totals; for instance, the daily sales data may be combined to produce monthly or yearly sales. This process is often used to build a data cube for data analysis at different levels of abstraction.
d. **Normalization**, where the attribute data is resized to fit a narrower range: −1.0 to 1.0; or 0.0 to 1.0.
e. **Discretization**, where interval labels replace the raw values of a numeric attribute (e.g., age) (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., *youth, adult*, *senior*). A concept hierarchy for the number attribute can then be created by recursively organizing the labels into higher-level concepts. To meet the demands of different users, more than one concept hierarchy might be built for the same characteristic.
f. **Concept hierarchy creation using nominal data** allows for the extrapolation of higher-level concepts like a street to concepts like a city or country. At the schema definition level, numerous hierarchies for nominal qualities can be automatically created and are implicit in the database structure.

Check Your Progress 1:

1. What is meant by data preprocessing?

2. Why is preprocessing important?
3. What are the 5 characteristics of data processing?
4. What are the 5 major steps of data preprocessing?
5. What is data cleaning?
6. What is the importance of data cleaning?
7. What are the main steps of Data Cleaning?

3.4 SELECTION AND DATA EXTRACTION

The process of choosing the best data source, data type, and collection tools is known as data selection. Prior to starting the actual data collection procedure, data selection is conducted. This concept makes a distinction between selective data reporting (excluding data that is not supportive of a study premise) and active/interactive data selection (using obtained data for monitoring activities/events or conducting secondary data analysis). Data integrity may be impacted by how acceptable data are selected for a research project.

The main goal of data selection is to choose the proper data type, source, and tool that enables researchers to effectively solve research issues. This decision typically depends on the discipline and is influenced by the research that has already been done, the body of that research, and the availability of the data sources.

Integrity issues may arise, when decisions about which "appropriate" data to collect, are primarily centred on cost and convenience considerations rather than the data's ability to successfully address research concerns. Cost and convenience are unquestionably important variables to consider while making a decision. However, researchers should consider how much these factors can skew the results of their study.

Data Selection Issues

When choosing data, researchers should be conscious of a few things, including:

- Researchers can appropriately respond to the stated research questions when the right type and appropriate data sources are used.
- Appropriate methods for obtaining a representative sample are used.
- The appropriate tools for gathering data are used. It is difficult to separate the choice of data type and source from the tools used to get the data. The type/source of data and the procedures used to collect it should be compatible.

Types and Sources of Data: Different data sources and types can be displayed in a variety of ways. There are two main categories of data:

- Quantitative data are expressed as numerical measurements at the interval

and ratio levels.
- Qualitative data can take the form of text, images, music, and video.

Although preferences within scientific disciplines differ as to which type of data is preferred, some researchers employ information from quantitative and qualitative sources to comprehend a certain event more thoroughly.

Researchers get data from people that may be qualitative (such as by studying child-rearing techniques) or quantitative (biochemical recording markers and anthropometric measurements). Field notes, journals, laboratory notes, specimens, and firsthand observations of people, animals, and plants can all be used as data sources. Data type and source interactions happen frequently.

Choosing the right data is discipline-specific and primarily influenced by the investigation's purpose, the body of prior research, and the availability of data sources. The following list of questions will help you choose the right data type and sources:

- What is the research question?
- What is the investigation's field of study? (This establishes the guidelines for any investigation. Selected data should not go beyond what is necessary for the investigation.
- According to the literature (prior research), what kind of information should be gathered?
- Which form of data—qualitative, quantitative, or a combination of both—should be considered?

Data extraction is the process of gathering or obtaining many forms of data from numerous sources, many of which may be erratically organized or wholly unstructured. Consolidating, processing and refining data enable information to be kept in a centralized area so that it can be altered. These venues could be local, online, or a combination of the two.

Data Extraction and ETL

To put the importance of data extraction into perspective, it is helpful to quickly assess the ETL process as a whole.

1. Extraction: One or more sources or systems are used to collect the data. Relevant data is located, identified, and then prepared for processing or transformation during the extraction phase. One can finally analyse them for business knowledge by combining various data types through extraction.
2. Transformation: It can be further refined after the data has been effectively extracted. Data is cleansed, sorted, and structured as part of the transformation process. Duplicate entries will be removed, missing values will be filled in or removed, and audits will be performed, for example, in order to offer data that is reliable, consistent, and usable.
3. Loading: The high-quality, converted data is subsequently sent to a single, centralized target location for storage and analysis.

Data extraction tools

The tools listed below can be used for tasks other than a simple extraction. They can be grouped into the following categories:

1. **Scrape storm:** One data extraction tool you may consider is the scrape

storm. It is software that uses AI to scrape the web or gather data. It is compatible with Windows, Mac, or Linux operating systems and has a simple and straightforward visual operation. This program automatically detects objects like emails, numbers, lists, forms, links, photos, and prices. When transferring extracted data to Excel, MongoDB, CSV, HTML, TXT, MySQL, SQL Server, PostgreSQL, Google Sheets, or WordPress, it can make use of a variety of export strategies.

2. **Altair Monarch**: Monarch is desktop-based, self-service, and does not require any programming. It can link to various data sources, including big data, cloud-based data, and both organized and unstructured data. It connects to, cleans, and processes data with high speed and no errors. It employs more than 80 built-in data preparation functions. Less time is wasted on making data legible so that more time can be spent on creating higher-level knowledge.
3. **Klippa:** The processing of contracts, invoices, receipts, and passports can be done using the cloud with Klippa. For the majority of documents, the conversion time may be between one and five seconds. The data classification and manipulation may be done online, round-the-clock, and supports a variety of file types, including PDF, JPG, and PNG. It can also convert between JSON, PDF/A, XLSX, CSV, and XML. Additionally, the software handles file sharing, custom branding, payment processing, cost management, and invoicing management.
4. NodeXL: For Microsoft Excel 2007, 2010, 2013, and 2016, Basic is a free, open-source add-on extension. Since the software is an add-on, data integration is not performed; instead, it focuses on social network analytics. Advanced network analytics, text and sentiment analysis, and robust report generating are extra capabilities included with NodeXL Pro.

Check Your Progress 2:

1. What is the data selection process??
2. What is Data Extraction and define the term ETL?
3. What are the challenges of data extraction?

3.5 DATA CURATION

Data curation is creating, organizing and managing data sets so that people looking for information can access and use them. It comprises compiling, arranging, indexing, and categorizing data for users inside of a company, a group, or the general public. To support business decisions, academic needs, scientific research, and other initiatives, data can be curated. Data curation is a step in the larger data management process that helps prepare data sets for usage in business intelligence (BI) and analytics applications. In other cases, the curation process might be fed with ready-made data for ongoing management and maintenance. In organizations without particular data curator employment, data stewards, data engineers, database administrators, data scientists, or business users may fill that role.

3.5.1 Steps of data curation

There are numerous tasks involved in curating data sets, which can be divided into the following main steps.

- The data that will be required for the proposed analytics applications should be determined.
- Map the data sets and note the metadata that goes with them.
- Collect the data sets.
- The data should be ingested into a system, a data lake, a data warehouse, etc.
- Cleanse the data to remove abnormalities, inconsistencies, and mistakes, including missing values, duplicate records, and spelling mistakes.
- Model, organize, and transform the data to prepare it for specific analytics applications.
- To make the data sets accessible to users, create searchable indexes of them.
- Maintain and manage the data in compliance with the requirements of continuous analytics and the laws governing data privacy and security.

3.5.2 Importance of Data Curation

The following are the reasons for performing data curation.

1. Helps to organize pre-existing data for a corporation: Businesses produce a large amount of data on a regular basis, however, this data can occasionally be lacking. When a customer clicks on a website, adds something to their cart, or completes a transaction, an online clothes retailer might record that information. Data curators assist businesses in better understanding vast amounts of information by assembling prior data into data sets.
2. Connects professionals in different departments: When a company engages in data curation, it often brings together people from several departments who might not typically collaborate. Data curators might collaborate with stakeholders, system designers, data scientists, and data analysts to collect and transfer information.
3. High-quality data typically uses organizational techniques that make it simple to grasp and have fewer errors. Curators may make sure that a company's research and information continue to be of the highest caliber because the data curation process entails cleansing the data. Removing unnecessary information makes research more concise, which may facilitate better data set structure.
4. Makes data easy to understand: Data curators make sure there are no errors and utilize proper formatting. This makes it simpler for specialists who are not knowledgeable about a research issue to comprehend a data set.
5. Allows for higher cost and time efficiency: A business may spend more time and money organizing and distributing data if it does not regularly employ data curation. Because prior data is already organized and distributed, businesses that routinely do data curation may be able to save

time, effort, and money. Businesses can reduce the time it takes to obtain and process data by using data curators, who handle the data.

Check Your Progress 3:

1. What is the Importance of Data Curation?
2. Explain Data Curation.
3. What are the goals of data curation?
4. What are the benefits of data curation?

3.6 DATA INTEGRATION

Data integration creates coherent data storage by combining data from several sources. Smooth data integration is facilitated by the resolution of semantic heterogeneity, metadata, correlation analysis, tuple duplicate identification, and data conflict detection. It is a tactic that combines data from several sources so that consumers may access it in a single, consistent view that displays their status. Systems can communicate using flat files, data cubes, or numerous databases. Data integration is crucial because it maintains data accuracy while providing a consistent view of dispersed data. It helps the analysis tools extract valuable information, which in turn helps the executive and management make tactical choices that will benefit the company.

3.6.1 Data Integration Techniques

Manual Integration-When integrating data, this technique avoids employing automation. The data analyst gathers, purifies, and integrates the data to create actionable data. A small business with a modest amount of data can use this approach. Nevertheless, the extensive, complex, and ongoing integration will take a lot of time. It takes time because every step of the process must be performed manually.

Middleware Integration-Data from many sources are combined, normalized, and stored in the final data set using middleware software. This method is employed when an organization has to integrate data from historical systems into modern systems. Software called middleware serves as a translator between antiquated and modern systems. You could bring an adapter that enables the connection of two systems with various interfaces. It only works with specific systems.

Application-based integration- To extract, transform, and load data from various sources, it uses software applications. Although this strategy saves time and effort, it is a little more difficult because creating such an application requires technical knowledge.

Uniform Access Integration- This method integrates information from a wider range of sources. In this instance, however, the data is left in its initial place and is not moved. To put it simply, this technique produces a unified view of the combined data. The integrated data does not need to be saved separately because the end user only sees the integrated view.

3.6.2 Data Integration Approaches

There are two basic data integration approaches. These are –

Tight Coupling- It combines data from many sources into a single physical location using ETL (Extraction, Transformation, and Loading) tools.

Loose Coupling- The real source databases are the most efficient place to store facts with loose coupling. This method offers an interface that receives a user query, converts it into a format that the source databases can understand, and immediately transmits the question to the source databases to get the answer.

Check Your Progress 4:

1. What is meant by data integration?
2. What is an example of data integration?
3. What is the purpose of data integration?

3.7 KNOWLEDGE DISCOVERY

Knowledge discovery in databases is the process of obtaining pertinent knowledge from a body of data (KDD). This well-known knowledge discovery method includes several processes, such as data preparation and selection, data cleansing, incorporating prior knowledge about the data sets, and interpreting precise answers from the observed results.

Marketing, fraud detection, telecommunications, and manufacturing are some of the key KDD application areas. In the last ten years, the KDD process has reached its pinnacle. Inductive learning, Bayesian statistics, semantic query optimization, knowledge acquisition for expert systems, and information theory are just a few of the numerous discovery-related methodologies it now houses. Extraction of high-level knowledge from low-level data is the ultimate objective. Due to the accessibility and quantity of data available today, knowledge discovery is a challenge of astounding importance and necessity. Given how swiftly the topic has expanded, it is not surprising that professionals and experts today have access to a variety of treatments.

Steps of Knowledge Discovery

1. Developing an understanding of the application domain: Knowledge discovery starts with this preliminary step. It establishes the framework for selecting the best course of action for a variety of options, such as transformation, algorithms, representation, etc. The individuals in charge of a KDD project need to be aware of the end users' goals as well as the environment in which the knowledge discovery process will take place.

2. Selecting and producing the data set that will be used for discovery -Once the objectives have been specified, the data that will be used for the knowledge discovery process should be identified. Determining what data is accessible, obtaining essential information, and then combining all the data for knowledge discovery into one set are the factors that will be considered for the procedure. Knowledge discovery is important since it extracts knowledge and insight from the given data. This provides the framework for building the models.

3. Preprocessing and cleansing – This step helps in increasing the data reliability. It comprises data cleaning, like handling the missing quantities and removing noise or outliers. In this situation, it might make use of sophisticated statistical methods or an analysis algorithm. For instance, the goal of the Data Mining supervised approach may change if it is determined that a certain attribute is unreliable or has a sizable amount of missing data. After developing a prediction model for these features, missing data can be forecasted. A variety of factors affect how much attention is paid to this level. However, breaking down the components is important and frequently useful for enterprise data frameworks.

4. Data Transformation-This phase entails creating and getting ready the necessary data for knowledge discovery. Here, techniques of attribute transformation (such as discretization of numerical attributes and functional transformation) and dimension reduction (such as feature selection, feature extraction, record sampling etc.) are employed. This step, which is frequently very project-specific, can be important for the success of the KDD project. Proper transformation results in proper analysis and proper conclusions.

5. Prediction and description- The decisions to use classification, regression, clustering, or any other method can now be made. Mostly, this uses the KDD objectives and the decisions made in the earlier phases. A forecast and a description are two of the main objectives of knowledge discovery. The visualization aspects are included in descriptive knowledge discovery. Inductive learning, which generalizes a sufficient number of prepared models to produce a model either explicitly or implicitly, is used by the majority of knowledge discovery techniques. The fundamental premise of the inductive technique is that the prepared model holds true for the examples that follow.

6. Deciding on knowledge discovery algorithm -We now choose the strategies after determining the technique. In this step, a specific technique must be chosen to be applied while looking for patterns with numerous inducers. If precision and understandability are compared, the former is improved by neural networks, while decision trees improve the latter. There are numerous ways that each meta-learning system could be successful. The goal of meta-learning is to explain why a data analysis algorithm is successful or unsuccessful in solving a particular problem. As a result, this methodology seeks to comprehend the circumstances in which a data analysis algorithm is most effective. Every algorithm has parameters and learning techniques, including tenfold cross-validation or a different division for training and testing.

7. Utilizing the Data Analysis Algorithm-Finally, the data analysis algorithm is put into practice. The approach might need to be applied several times before producing a suitable outcome at this point. For instance, by rotating the algorithms, you can alter variables like the bare minimum of instances in a single decision tree leaf.

8. Evaluation-In this stage, the patterns, principles, and dependability of the results of the knowledge discovery process are assessed and interpreted in light of the objective outlined in the preceding step. Here, we take into account the preprocessing steps and how they impact the final results. As an illustration, add a feature in step 4 and then proceed. The primary considerations in this step are the understanding and utility of the induced model. In this stage, the identified knowledge is also documented for later use.

Check Your Progress 5:

1. What is Knowledge Discovery?
2. What are the Steps involved in Knowledge Discovery?
3. What are knowledge discovery tools?
4. Explain the process of KDD.

3.8 SUMMARY

Despite the development of several methods for preparing data, the intricacy of the issue and the vast amount of inconsistent or unclean data mean that this field of study is still very active. This unit gives a general overview of data pre-processing and describes how to turn raw data into usable information. The preprocessing of the raw data included data integration, data reduction, transformation, and discretization. In this unit, we have discussed five different data-cleaning techniques that can make data more reliable and produce high-quality results. Building, organizing, and maintaining data sets is known as data curation. A data curator usually determines the necessary data sets and makes sure they are gathered, cleaned up, and changed as necessary. The curator is also in charge of providing users with access to the data sets and information related to them, such as their metadata and lineage documentation. The primary goal of the data curator is to make sure users have access to the appropriate data for analysis and decision-making. Data integration is the procedure of fusing information from diverse sources into a single, coherent data store. The unit also introduced knowledge discovery techniques and procedures.

3.9 SOLUTIONS/ANSWERS

Check Your Progress 1:

1. As a part of data preparation, data preprocessing refers to any type of processing done on raw data to get it ready for a data processing technique. It has long been regarded as a crucial first stage in the data mining process.

2. It raises the reliability and accuracy of the data. Preprocessing data can increase the correctness and quality of a dataset, making it more dependable by removing missing or inconsistent data values brought by human or computer mistakes. It ensures consistency in data.

3. Data quality is characterized by five characteristics: correctness, completeness, reliability, relevance, and timeliness.

4. The five major steps of data preprocessing are:
   - Data quality assessment
   - Data cleaning
   - Data transformation
   - Data reduction

5. The practice of correcting or deleting inaccurate, damaged, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are numerous ways for data to be duplicated or incorrectly categorized when merging multiple data sources.

6. Data cleansing, sometimes referred to as data cleaning or scrubbing, is the process of locating and eliminating mistakes, duplication, and irrelevant data from a raw dataset. Data cleansing, which is a step in the preparation of data, ensures that the cleaned data is used to create accurate, tenable visualizations, models, and business choices.

7. Step 1: Remove irrelevant data; Step 2: Deduplicate your data; Step 3: Fix structural errors; Step 4: Deal with missing data; Step 5: Filter out data outliers; Step 6: Validate your data.

Check Your Progress 2:

1. The process of retrieving data from the database that are pertinent to the analysis activity is known as data selection. Sometimes the data selection process comes before data transformation and consolidation.
2. Data extraction is the process of gathering or obtaining many forms of data from numerous sources, many of which may be erratically organized or wholly unstructured. The process of extracting, transforming, and loading data is called ETL. Thus, ETL integrates information from several data sources into a single consistent data store, which can be a data warehouse or data analytics system.
3. The cost and time involved in extracting data, as well as the accuracy of the data, are obstacles. The correctness of the data depends on the quality of the data source, which can be an expensive and time-consuming procedure.

Check Your Progress 3:

1. It entails gathering, organizing, indexing, and cataloguing information for users within an organization, a group, or the wider public. Data curation can help with academic needs, commercial decisions, scientific research, and other endeavors.
2. The process of producing, arranging, and managing data sets so that users who are looking for information can access and use them is known as data curation. Data must be gathered, organized, indexed, and catalogued for users within an organization, group, or the broader public.
3. By gathering pertinent data into organized, searchable data assets, data curation's overarching goal is to speed up the process of extracting

insights from raw data.

4. The benefits of data curation are:
   - Easily discover and use data.
   - Ensure data quality.
   - Maintain metadata linked with data.
   - Ensure compliance through data lineage and classification.

Check Your Progress 4:

1. Data integration is used to bring together data from several sources to give people a single perspective. Making data more readily available, easier to consume, and easier to use by systems and users is the foundation of data integration.
2. In the case of customer data integration, information about each customer is extracted from several business systems, such as sales, accounts, and marketing, and combined into a single picture of the client for use in customer service, reporting, and analysis.
3. Data integration combines data collected from various platforms to increase its value for your company. It enables your staff to collaborate more effectively and provide more for your clients. You cannot access the data collected in different systems without data integration.

Check Your Progress 5:

1. Knowledge discovery is the labour-intensive process of extracting implicit, unknown-before information from databases that may be beneficial.
2. Steps of Knowledge Discovery:
   - Developing an understanding of the application domain
   - Selecting and producing the data set that will be used for the discovery
   - Preprocessing and cleansing
   - Data Transformation
   - Prediction and description
   - Deciding on a data analysis algorithm
   - Utilizing the data analysis algorithm
   - Evaluation
3. The process can benefit from a variety of qualitative and quantitative methods and techniques, such as knowledge surveys, questionnaires, one-on-one and group interviews, focus groups, network analysis, and observation. It can be used to locate communities and specialists.
4. Knowledge Discovery from Data, often known as KDD, is another commonly used phrase that is treated as a synonym for data mining. Others see data mining as just a crucial stage in the knowledge discovery process when intelligent techniques are used to extract data patterns. The steps involved in knowledge discovery from data are as follows: :
   - Data cleaning (to remove noise or irrelevant data).
   - Data integration (where multiple data sources may be combined).
   - Data selection (where data relevant to the analysis task are retrieved from the database).
   - Data transformation (where data are consolidated into forms appropriate for mining by performing summary or aggregation functions, for sample).

- Data mining (an important process where intelligent methods are applied in order to extract data patterns).
- Pattern evaluation (to identify the fascinating patterns representing knowledge based on some interestingness measures).
- Knowledge presentation (where knowledge representation and visualization techniques are used to present the mined knowledge to the user).

3.10 FURTHER READINGS

References

Data Preprocessing in Data Mining - GeeksforGeeks. (2019, March 12). GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Data Cleaning in Data Mining - Javatpoint. (n.d.). Www.Javatpoint.Com. Retrieved February 11, 2023, from https://www.javatpoint.com/data-cleaning-in-data-mining

Data Integration in Data Mining - GeeksforGeeks. (2019, June 27). GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/data-integration-in-data-mining/

Dowd, R., Recker, R.R., Heaney, R.P. (2000). Study subjects and ordinary patients. Osteoporos Int. 11(6): 533-6.

Fourcroy, J.L. (1994). Women and the development of drugs: why can't a woman be more like a man? Ann N Y Acad Sci, 736:174-95.

Goehring, C., Perrier, A., Morabia, A. (2004). Spectrum Bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. Statistics in Medicine, 23(1):125-35.

Gurwitz,J.H., Col. N.F., Avorn, J. (1992). The exclusion of the elderly and women from clinical trials in acute myocardial infarction. JAMA, 268(11): 1417-22.

Hartt, J., Waller, G. (2002). Child abuse, dissociation, and core beliefs in bulimic disorders. Child Abuse Negl. 26(9): 923-38.

Kahn, K.S, Khan, S.F, Nwosu, C.R, Arnott, N, Chien, P.F.(1999). Misleading authors' inferences in obstetric diagnostic test literature. American Journal of Obstetrics and Gynaecology., 181(1`), 112-5.

KDD Process in Data Mining - GeeksforGeeks. (2018, June 11). GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/kdd-process-in-data-mining/

Maynard, C., Selker, H.P., Beshansky, J.R.., Griffith, J.L., Schmid, C.H., Califf, R.M., D'Agostino, R.B., Laks, M.M., Lee, K.L., Wagner, G.S., et al. (1995). The exclusions of women from clinical trials of thrombolytic therapy: implications for developing the thrombolytic predictive instrument database. Med Decis Making (Medical Decision making: an international journal of the Society for Medical Decision Making), 15(1): 38-43.

Pratt, M. K. (2022, January 31). What is Data Curation? - Definition from SearchBusinessAnalytics. Business Analytics; TechTarget. https://www.techtarget.com/searchbusinessanalytics/definition/data-curation

Robinson, D., Woerner, M.G., Pollack, S., Lerner, G. (1996). Subject selection bias in clinical: data from a multicenter schizophrenia treatment center. Journal of Clinical Psychopharmacology, 16(2): 170-6.

Sharpe, N. (2002). Clinical trials and the real world: selection bias and generalisability of trial results. Cardiovascular Drugs and Therapy, 16(1): 75-7.

Walter, S.D., Irwig, L., Glasziou, P.P. (1999). Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol., 52(10): 943-51.

What is Data Extraction? Definition and Examples | Talend. (n.d.). Talend - A Leader in Data Integration & Data Integrity. Retrieved February 11, 2023, from https://www.talend.com/resources/data-extraction-defined/

Whitney, C.W., Lind, B.K., Wahl, P.W. (1998). Quality assurance and quality control in longitudinal studies. Epidemiologic Reviews, 20(1): 71-80.