- **Data: It is how the data objects and their attributes are stored.**
- **An attribute is an object's property or characteristics. For example. A person's hair colour, air humidity etc.**
- **An attribute set defines an object. The object is also referred to as a record of the instances or entity.**
- **Different types of attributes or data types:**
- **In data mining, understanding the different types of attributes or data types is essential as it helps to determine the appropriate data analysis techniques to use. The following are the different types of data:**
- **1]Nominal Data:**
- **This type of data is also referred to as categorical data. Nominal data represents data that is qualitative and cannot be measured or compared with numbers. In nominal data, the values represent a category, and there is no inherent order or hierarchy. Examples of nominal data include gender, race, religion, and occupation. Nominal data is used in data mining for classification and clustering tasks.**

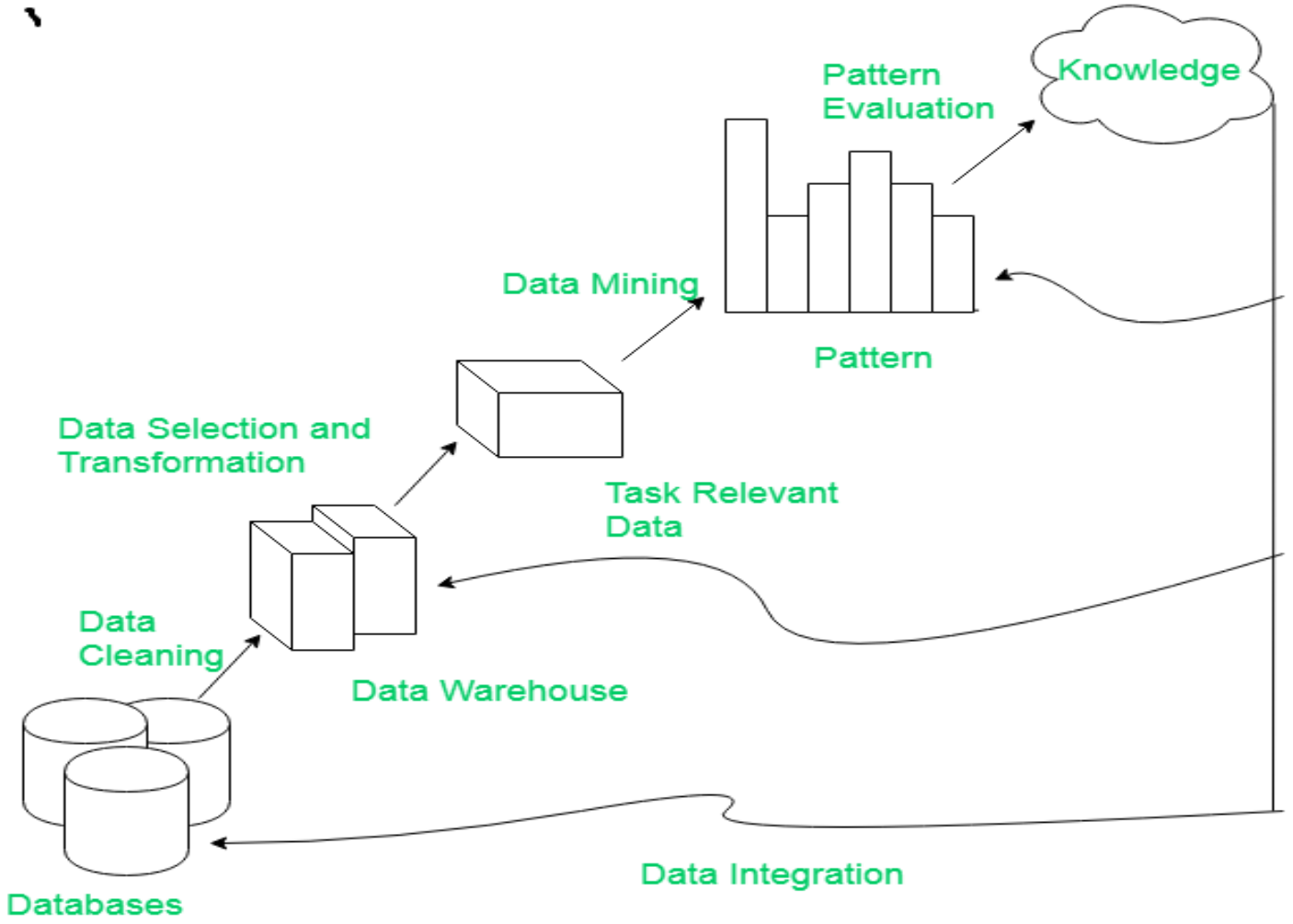
- **2]Ordinal Data:**
- **This type of data is also categorical, but with an inherent order or hierarchy. Ordinal data represents qualitative data that can be ranked in a particular order. For instance, education level can be ranked from primary to tertiary, and social status can be ranked from low to high. In ordinal data, the distance between values is not uniform. This means that it is not possible to say that the difference between high and medium social status is the same as the difference between medium and low social status. Ordinal data is used in data mining for ranking and classification tasks.**
- **3]Binary Data:**
- **This type of data has only two possible values, often represented as 0 or 1. Binary data is commonly used in classification tasks, where the target variable has only two possible outcomes. Examples of binary data include yes/no, true/false, and pass/fail. Binary data is used in data mining for classification and association rule mining tasks.**

- **4]Interval Data:**
- This type of data represents quantitative data with equal intervals between consecutive values. Interval data has no absolute zero point, and therefore, ratios cannot be computed. Examples of interval data include temperature, IQ scores, and time. Interval data is used in data mining for clustering and prediction tasks.
- **5]Ratio Data:**
- This type of data is similar to interval data, but with an absolute zero point. In ratio data, it is possible to compute ratios of two values, and this makes it possible to make meaningful comparisons. Examples of ratio data include height, weight, and income. Ratio data is used in data mining for prediction and association rule mining tasks.
- **6]Text Data:**
- This type of data represents unstructured data in the form of text. Text data can be found in social media posts, customer reviews, and news articles. Text data is used in data mining for sentiment analysis, text classification, and topic modeling tasks.
- dataset. Data may be in different formats, such as text, numerical, or categorical. Preprocessing techniques, such as data transformation and normalization, can be used to convert data into a consistent format for analysis.

- **KDD Process in Data Mining**
- The need of data mining is to extract useful information from large datasets and use it to make predictions or better decision-making. Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.
- For examples: Banking sector, Market Basket Analysis, Network Intrusion Detection.
-

- **KDD Process**
- KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.The following steps are included in KDD process:

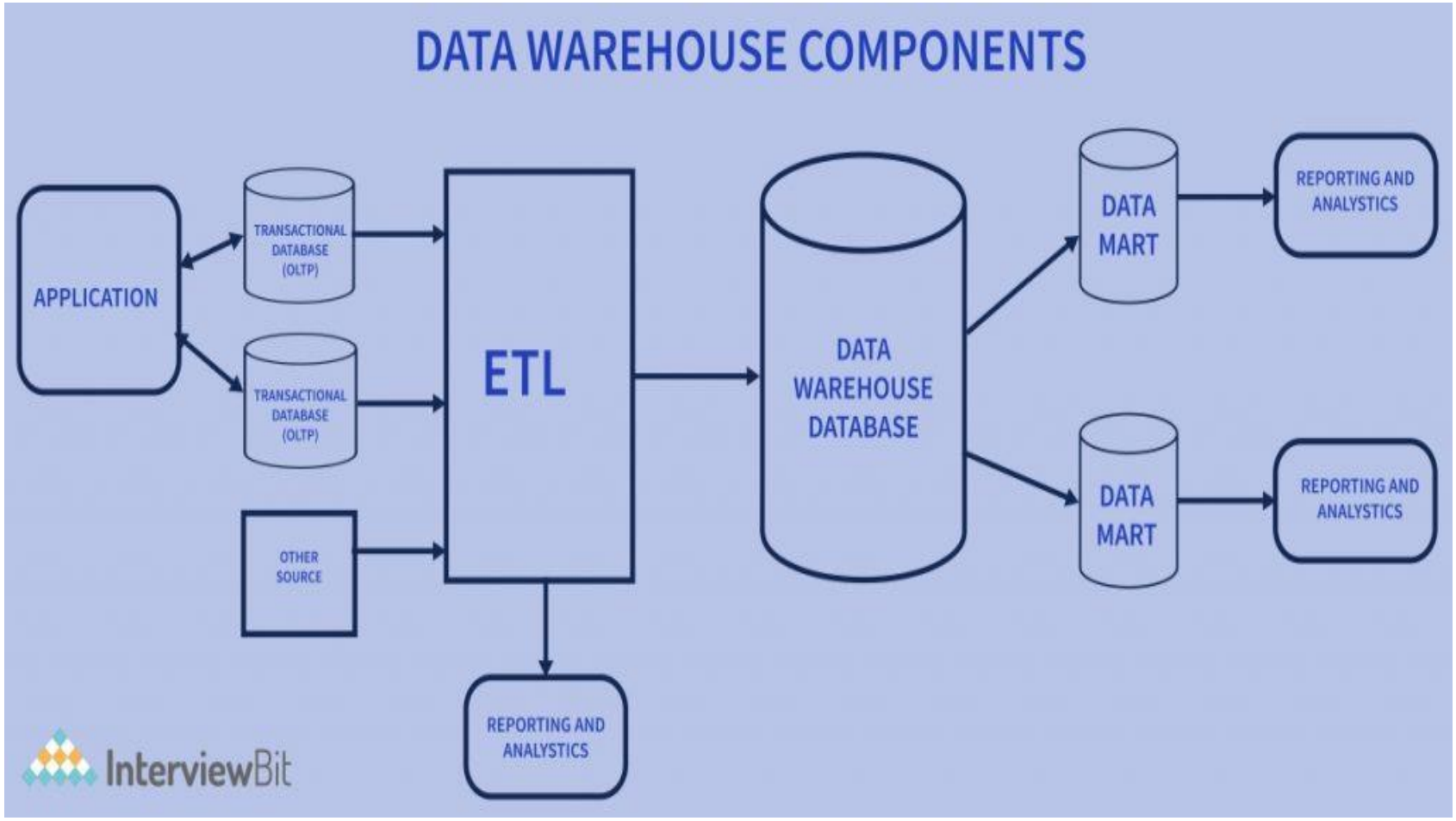
KDD is an **iterative process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.**Preprocessing of databases** consists of **Data cleaning** and **Data Integration**.


Knowledge
Pattern Evaluation
Data Mining
Pattern
Data Selection and Transformation
Task Relevant Data
Data Cleaning
Data Warehouse
Databases
Data Integration

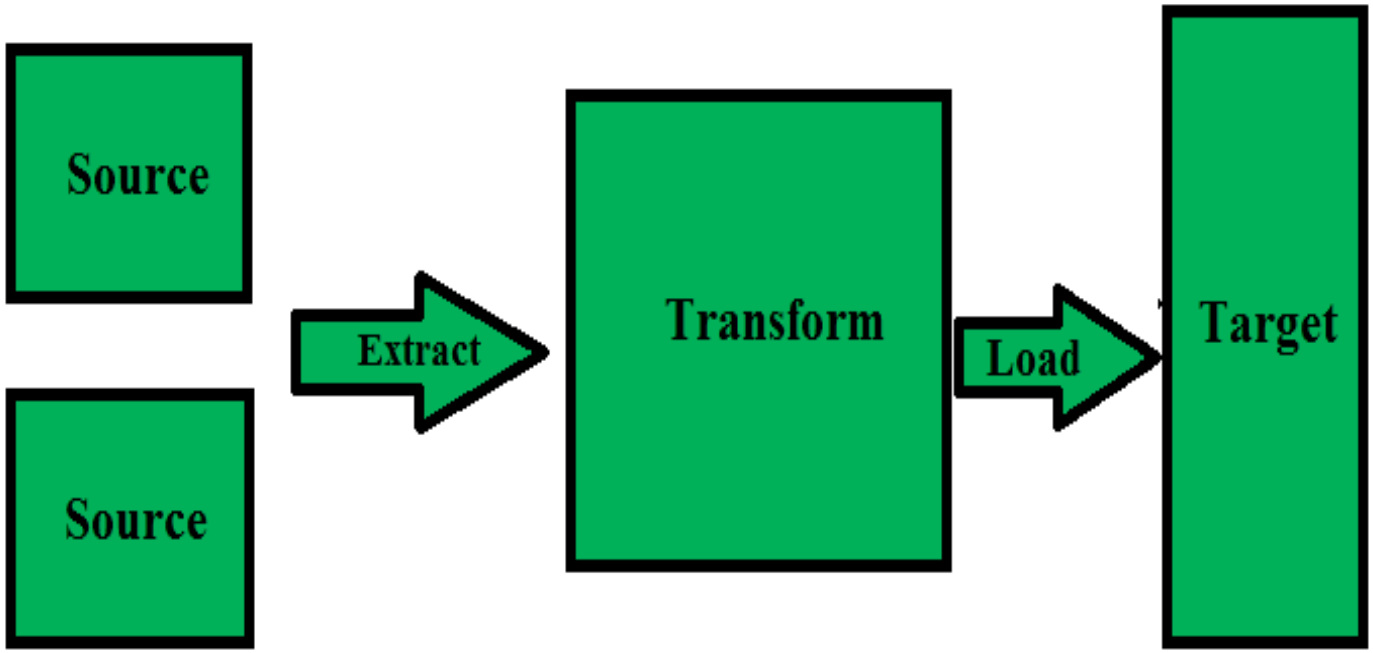
- **Data Cleaning**
- **Data cleaning is defined as removal of noisy and irrelevant data from collection.**
- **Cleaning in case of Missing values.**
- **Cleaning noisy data, where noise is a random or variance error.**
- **Cleaning with Data discrepancy detection and Data transformation tools.**
- **Data Integration**
- **Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse). Data integration using Data Migration tools, Data Synchronization tools and ETL(Extract-Load-Transformation) process.**
- **Data Selection**
- **Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use Neural network, Decision Trees, Naive bayes, Clustering, and Regression methods.**

- Data **Transformation**
- **Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:**
- **Data Mapping: Assigning elements from source base to destination to capture transformations.**
- **Code generation: Creation of the actual transformation program.**
- **Data Mining**
- **Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into patterns, and decides purpose of model using classification or characterization.**
- **Pattern Evaluation**
- **Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find interestingness score of each pattern, and uses summarization and Visualization to make data understandable by user.**
- **Knowledge Representation**
- **This involves presenting the results in a way that is meaningful and can be used to make decisions.**

S.No.	Database	Data Warehouse
1.	A common Database is based on operational or transactional processing. Each operation is an indivisible transaction.	A Data Warehouse is based on analytical processing.
2.	Generally, a Database stores current and up-to-date data which is used for daily operations.	A Data Warehouse maintains historical data over time. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.
3.	A database is generally application specific. Example - A database stores related data, such as the student details in a school.	A Data Warehouse is integrated generally at the organization level, by combining data from different databases. Example - A data warehouse integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city.
4.	Constructing a Database is not so expensive.	Constructing a Data Warehouse can be expensive.


DATA WAREHOUSE COMPONENTS
APPLICATION
TRANSACTIONAL DATABASE (OLTP)
TRANSACTIONAL DATABASE (OLTP)
OTHER SOURCE
ETL
DATA WAREHOUSE DATABASE
DATA MART
DATA MART
REPORTING AND ANALYSTICS
REPORTING AND ANALYSTICS
REPORTING AND ANALYSTICS
InterviewBit

- **Data Warehousing:**
- It is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed rather than transaction processing.
- A **data warehouse** is designed to support the management decision-making process by providing a platform for data cleaning, data integration, and data consolidation.
- A data warehouse contains subject-oriented, integrated, time-variant, and non-volatile data. The Data warehouse consolidates data from many sources while ensuring data quality, consistency, and accuracy.
- Data warehouse improves system performance by separating analytics processing from transactional databases. Data flows into a data warehouse from the various databases. A data warehouse works by organizing data into a schema that describes the layout and type of data. Query tools analyze the data tables using schema.
-


Source
Source
Extract
Transform
Load
Target

- **Advantages of Data Warehousing:**
- The data warehouse’s job is to make any form of corporate data easier to understand. The majority of the user’s job will consist of inputting raw data.
- The capacity to update continuously and frequently is the key benefit of this technology. As a result, data warehouses are perfect for organizations and entrepreneurs who want to stay current with their target audience and customers.
- It makes data more accessible to businesses and organizations.
- A data warehouse holds a large volume of historical data that users can use to evaluate different periods and trends in order to create predictions for the future.
- **Disadvantages of Data Warehousing:**
- There is a great risk of accumulating irrelevant and useless data. Data loss and erasure are other potential issues.
- Data is gathered from various sources in a data warehouse. Cleansing and transformation of the data are required. This could be a difficult task.

Database System	Data Warehouse
It supports operational processes.	It supports analysis and performance reporting.
Capture and maintain the data.	Explore the data.
Current data.	Multiple years of history.
Data is balanced within the scope of this one system.	Data must be integrated and balanced from multiple system.
Data is updated when transaction occurs.	Data is updated on scheduled processes.
Data verification occurs when entry is done.	Data verification occurs after the fact.
100 MB to GB.	100 GB to TB.
ER based.	Star/Snowflake.
Application oriented.	Subject oriented.
Primitive and highly detailed.	Summarized and consolidated.
Flat relational.	Multidimensional.