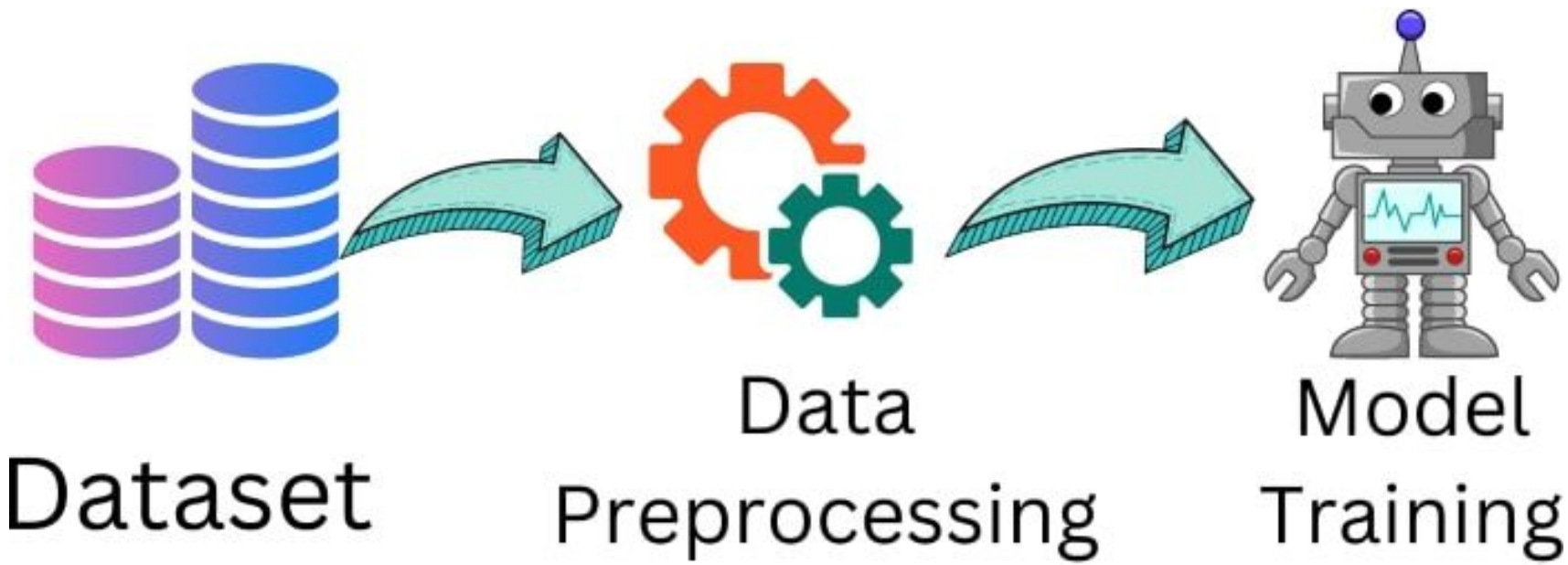
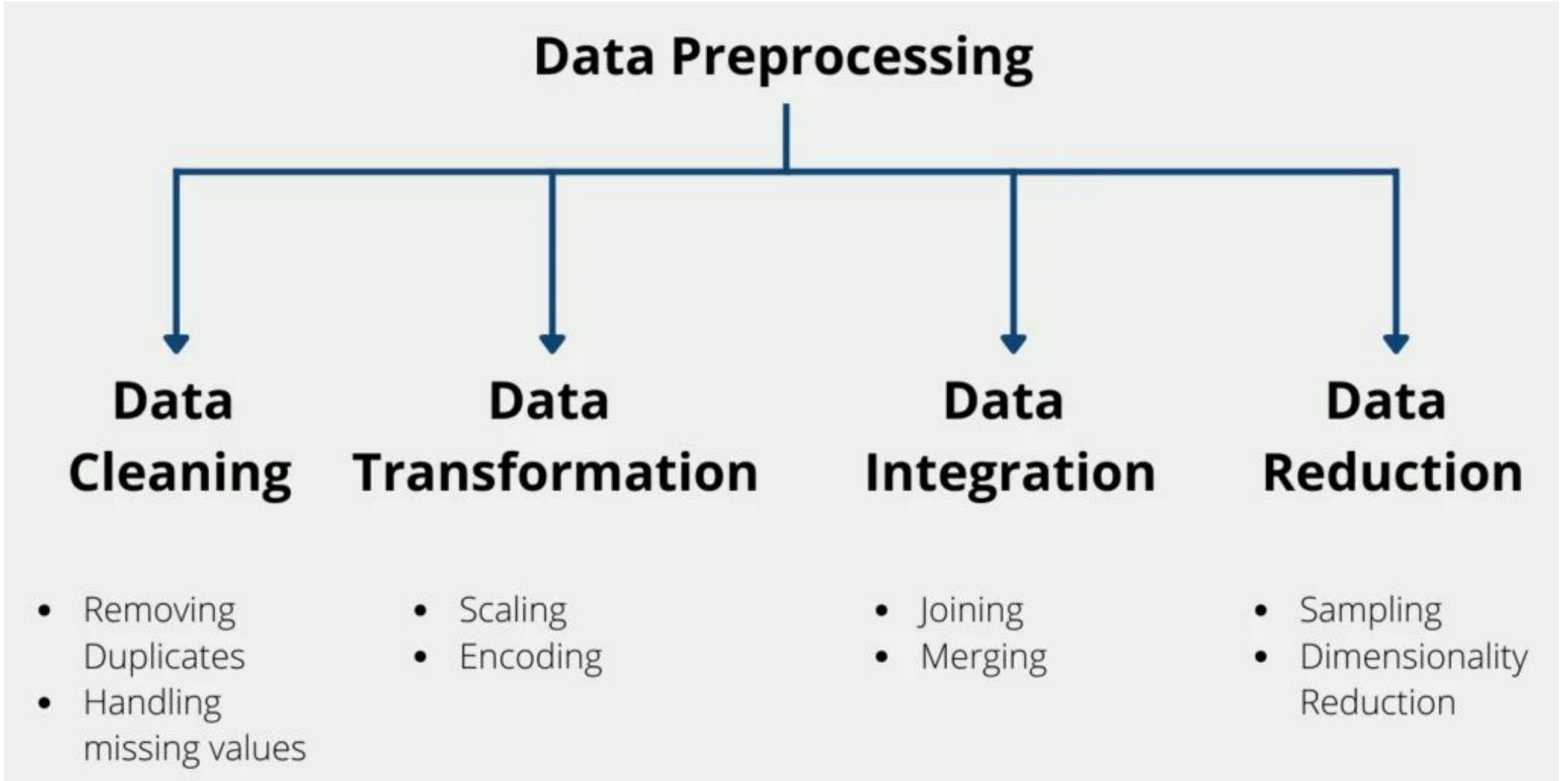
Data Preprocessing



What is data preprocessing?

- Data preprocessing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure.
- Data preprocessing techniques have been adapted for training machine learning models and AI models and for running inferences against them.
- Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.


Data Preprocessing
Data Cleaning
Data Transformation
Data Integration
Data Reduction
• Removing Duplicates
• Handling missing values
• Scaling
• Encoding
• Joining
• Merging
• Sampling
• Dimensionality Reduction

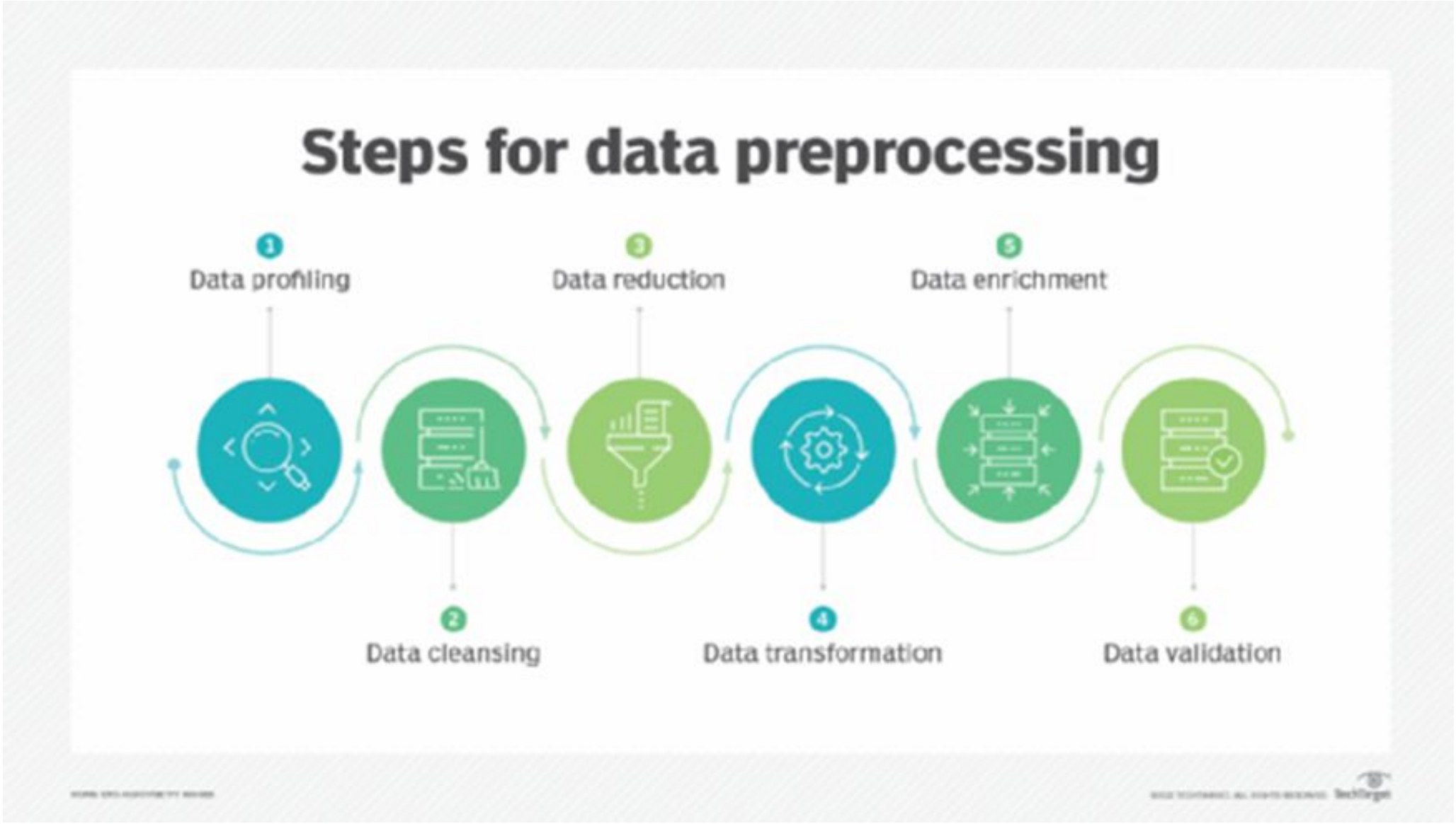
There are several different tools and methods used for preprocessing data, including the following:

- **Sampling**:which selects a representative subset from a large population of data.
- **Transformation**: which manipulates raw data to produce a single input.
- **Denoising**: which removes noise from data. It can occur due to data entry errors, faulty data collection, etc
- **Imputation**: which synthesizes statistically relevant data for missing values(substituting missing **data** with a different value).
- **Normalization**: which organizes data for more efficient access; and
- **Feature extraction**, which pulls out a relevant feature subset that is significant in a particular context.

Why is data preprocessing important?

- Data Preprocessing is an important step in the Data Preparation stage of a Data Science development lifecycle that will ensure reliable, robust, and consistent results. The main objective of this step is to ensure and check the quality of data before applying any Machine Learning or Data Mining methods. Let's review some of its benefits -
- **Accuracy** - Data Preprocessing will ensure that input data is accurate and reliable by ensuring there are no manual entry errors, no duplicates, etc.
- **Completeness** - It ensures that missing values are handled, and data is complete for further analysis.
- **Consistent** - Data Preprocessing ensures that input data is consistent, i.e., the same data kept in different places should match.
- **Timeliness** - Whether data is updated regularly and on a timely basis or not.
- **Trustable** - Whether data is coming from trustworthy sources or not.
- **Interpretability** - Raw data is generally unusable, and Data Preprocessing converts raw data into an interpretable format.

Key Steps in Data Preprocessing



Data profiling

Data profiling is the process of examining, analyzing and reviewing data to collect statistics about its quality. It starts with a survey of existing data and its characteristics.

Data scientists identify data sets that are pertinent to the problem at hand, inventory its significant attributes, and form a hypothesis of features that might be relevant for the proposed analytics or machine learning task. They also relate data sources to the relevant business concepts and consider which preprocessing libraries could be used.

Data Cleaning

- Data Cleaning uses methods to handle incorrect, incomplete, inconsistent, or missing values. Some of the techniques for Data Cleaning include -

Handling Missing Values

- Input data can contain missing or NULL values, which must be handled before applying any Machine Learning or Data Mining techniques.
- Missing values can be handled by many techniques, such as removing rows/columns containing NULL values and imputing NULL values using mean, mode, regression, etc.

De-noising

- De-noising is a process of removing noise from the data. Noisy data is meaningless data that is not interpretable or understandable by machines or humans. It can occur due to data entry errors, faulty data collection, etc.
- De-noising can be performed by applying many techniques, such as binning the features, using regression to smoothen the features to reduce noise, clustering to detect the outliers, etc.

Data Reduction

- Data Reduction is used to reduce the volume or size of the input data.
- Its main objective is to reduce storage and analysis costs and improve storage efficiency.

A few of the popular techniques to perform Data Reduction include –

- **Dimensionality Reduction** - It is the process of reducing the number of features in the input dataset. It can be performed in various ways, such as selecting features with the highest importance, Principal Component Analysis (PCA), etc.
- **Numerosity Reduction** - In this method, various techniques can be applied to reduce the volume of data by choosing alternative smaller representations of the data. For example, a variable can be approximated by a regression model, and instead of storing the entire variable, we can store the regression model to approximate it.
- **Data Compression** - In this method, data is compressed. Data Compression can be lossless or lossy depending on whether the information is lost or not during compression.

Data Transformation

- Data Transformation is a process of converting data into a format that helps in building efficient ML models and deriving better insights.

A few of the most common methods for Data Transformation include -

- **Smoothing** - Data Smoothing is used to remove noise in the dataset, and it helps identify important features and detect patterns. Therefore, it can help in predicting trends or future events.
- **Aggregation** - Data Aggregation is the process of **transforming large volumes of data into an organized and summarized format that is more understandable and comprehensive**. For example, a company may look at monthly sales data of a product instead of raw sales data to understand its performance better and forecast future sales.
- **Discretization** - Data Discretization is a process of converting numerical or continuous variables into a set of intervals/bins. This makes data easier to analyze. For example, the age features can be converted into various intervals such as (0-10, 11-20, ..) or (child, young, …).
- **Normalization** - Data Normalization is a process of converting a numeric variable into a specified range such as [-1,1], [0,1], etc. A few of the most common approaches to performing normalization are **Min-Max Normalization, Data Standardization or Data Scaling**, etc.

Data enrichment.

- Data enrichment is one of the key processes by which you can add more value to your data. It refines, improves and enhances your data set with the addition of new attributes.
- For example, using an address post code/ZIP field, you can take simple address data and enrich it by adding socio economic demographic data, such as average income, household size and population attributes. By enriching data in this way, you can get a better understanding of your customer base, and potential target customers.

Enrichment techniques

There are 6 common tasks involved in data enrichment:

- Appending Data
- Segmentation
- Derived Attributes
- Imputation
- Entity Extraction
- Categorization

Data validation.

- At this stage, the data is split into two sets.
- The first set is used to train a machine learning or deep learning model.
- The second set is the testing data that is used to gauge the accuracy and robustness of the resulting model.
- This second step helps identify any problems in the hypothesis used in the cleaning and feature engineering of the data. If the data scientists are satisfied with the results, they can push the preprocessing task to a data engineer who figures out how to scale it for production. If not, the data scientists can go back and make changes to the way they implemented the data cleansing and feature engineering steps.

How is data preprocessing used?

Data preprocessing plays a key role in earlier stages of machine learning and AI application development.

- In an AI context, data preprocessing is used to improve the way data is cleansed, transformed and structured to improve the accuracy of a new model, while reducing the amount of compute required.
- A good data preprocessing pipeline can create reusable components that make it easier to test out various ideas for streamlining business processes or improving customer satisfaction.
  - For example, preprocessing can improve the way data is organized for a recommendation engine by improving the age ranges used for categorizing customers.
- Preprocessing can also simplify the work of creating and modifying data for more accurate and targeted business intelligence insights.
  - For example, customers of different sizes, categories or regions may exhibit different behaviors across regions. Preprocessing the data into the appropriate forms could help BI teams weave these insights into BI dashboards.

Applications of Data Preprocessing

Data Preprocessing is important in the early stages of a Machine Learning and AI application development lifecycle.

A few of the most common usage or application include -

- **Improved Accuracy of ML Models** - Various techniques used to preprocess data, such as Data Cleaning, Transformation ensure that data is complete, accurate, and understandable, resulting in efficient and accurate ML models.
- **Reduced Costs** - Data Reduction techniques can help companies save storage and compute costs by reducing the volume of the data
- **Visualization** - Preprocessed data is easily consumable and understandable that can be further used to build dashboards to gain valuable insights.