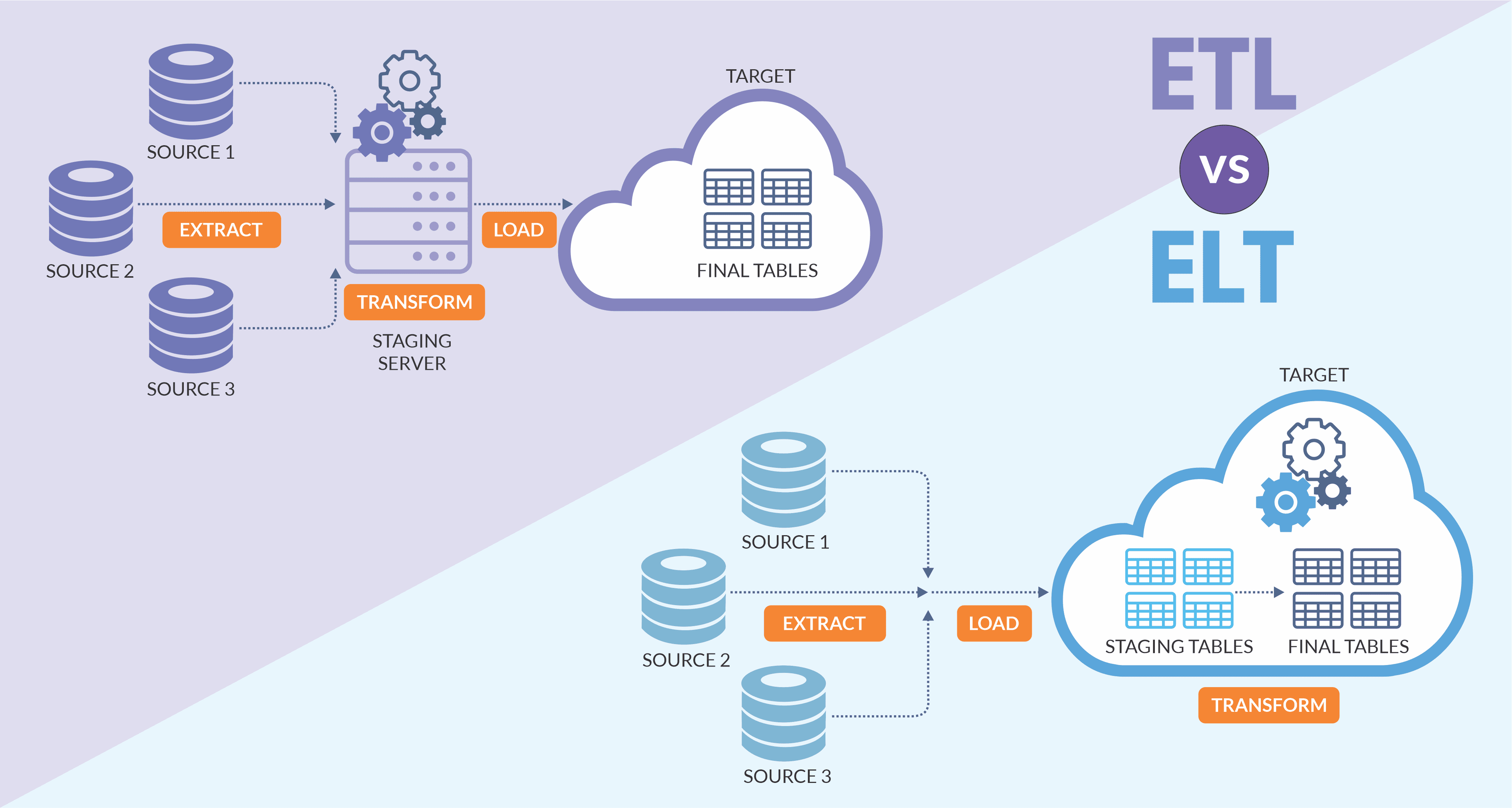
1.ETL & ELT

**What is ELT (Extract Load Transform)?**

ELT or Extract Load Transform is a process critical to the world of data integration and analytics. ELT refers to the process of [extracting data](https://bryteflow.com/data-extraction-for-etl-simplified/) from multiple sources, loading it to a target data warehouse and then using the resources of the data warehouse to power the transformation to make the data ready for consumption. When you ELT data you are taking advantage of a streamlined [ELT pipeline](https://bryteflow.com/data-pipelines-etl-pipelines-and-6-reasons-for-automation/) that can handle the tasks of extraction, loading and transformation to the data warehouse in real-time and in an automated manner. [ETL / ELT in Snowflake](https://bryteflow.com/snowflake_etl_snowflake_elt/)

**What is ETL (Extract Transform Load)?**

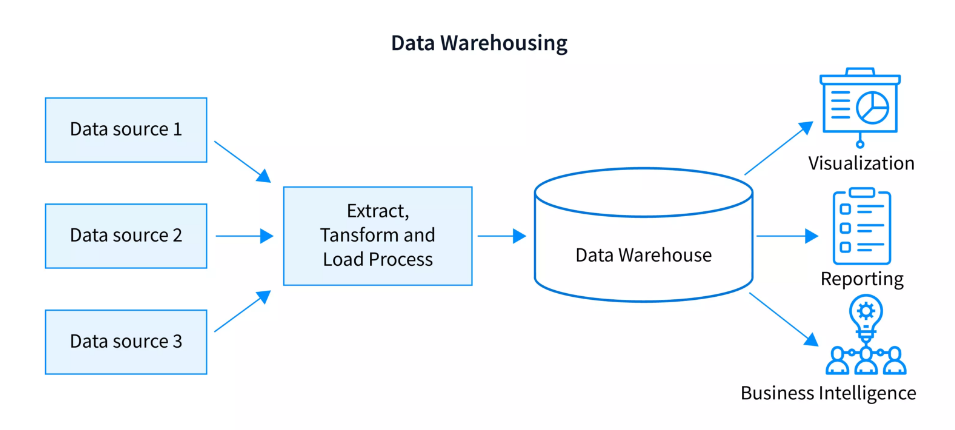
ETL or Extract Transform Load process refers to the 3 distinct steps: Extract – extracting and integrating data from different sources, Transform – holding and transforming it in a storage / staging area and Load – loading the transformed data to a data warehouse or data lake for further use. Data migrations and cloud data integrations require ETL to bring data across. ETL is an older process that came into being in the 70s while ELT came into being somewhere in the 2000s.

****

2.Data Warehouse Architecture

A Data Warehouse therefore can be described as a system that consolidates and manages data from different sources to assist an organization in making proper decisions. This makes the work of handling data to report easier. Two main construction approaches are used: Two of the most common models that have been developed are the Top-Down approach and the Bottom-Up approach and each of them possesses its strengths and weaknesses.

A D**ata-Warehouse** is a heterogeneous collection of data sources organized under a unified schema. There are 2 approaches for constructing a data warehouse: The top-down approach and the Bottom-up approach are explained below.



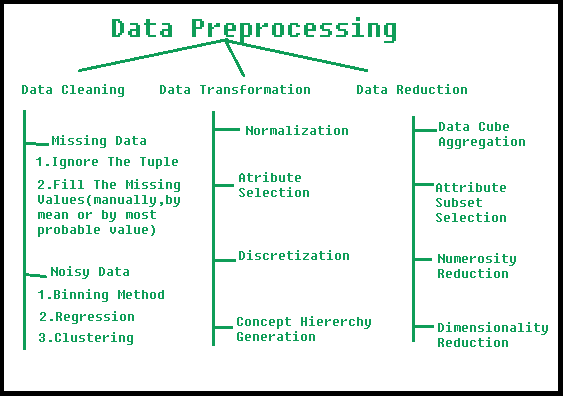
3.Data Preprocessing

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

**Steps of Data Preprocessing**

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

1. **Data Cleaning:**This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
2. **Data Integration:**This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.
3. **Data Transformation:**This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.
4. **Data Reduction:**This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.
5. **Data Discretization:**This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.
6. **Data Normalization:**This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.



4.How Neural Networks Can Be Used For Data Mining?

As all of us are aware that how technology is growing day-by-day and a Large amount of data is produced every second, analysing data is going to be very important because it helps us in fraud detection, identifying spam e-mail, etc.

So [Data Mining](https://www.geeksforgeeks.org/data-mining-techniques/) comes into existence to help us find hidden patterns, discover knowledge from large datasets.

Neural Network:

Neural Network is an information processing paradigm that is inspired by the human nervous system. As in the Human Nervous system, we have biological neurons in the same way in [Neural networks](https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/) we have Artificial Neurons which is a Mathematical Function that originates from biological neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. Each neuron receives signals through synapses that control the effects of the signal on the neuron.

5.Support Vector Machine (SVM) Algorithm

A Support Vector Machine (SVM) is a powerful machine learning algorithm widely used for both linear and nonlinear classification, as well as regression and outlier detection tasks. SVMs are highly adaptable, making them suitable for various applications such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection.

SVMs are particularly effective because they focus on finding the maximum separating hyperplane between the different classes in the target feature, making them robust for both binary and multiclass classification. In this outline, we will explore the Support Vector Machine (SVM) algorithm, its applications, and how it effectively handles both linear and nonlinear classification, as well as regression and outlier detection tasks.



6.KDD

KDD stands for Knowledge Discovery in Database

**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:

**Data Cleaning**

Data cleaning is defined as removal of noisy and irrelevant data from collection.

1. Cleaning in case of **Missing values**.
2. Cleaning **noisy** data, where noise is a random or variance error.
3. 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:

1. **Data Mapping**: Assigning elements from source base to destination to capture transformations.
2. **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 d**ecides 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.

KDD process

7.Agglomerative Clustering in Data Mining

Agglomerative clustering is a type of **hierarchical clustering**, where each data point starts as its own cluster, and over iterative steps, the algorithm merges the closest clusters until a stopping criterion is met (such as a desired number of clusters or a distance threshold). This is a **bottom-up approach**, in contrast to divisive clustering, which works top-down by starting with one cluster and splitting it.

**1. How Agglomerative Clustering Works:**

The process involves several steps:

1. **Start with each data point as an individual cluster**.
2. **Merge the two closest clusters** based on a chosen distance metric.
3. **Repeat step 2** until only one cluster remains or a stopping criterion is met.

**2. Key Concepts:**

* **Dendrogram**: A tree-like diagram that records the sequences of merges or splits. It helps visualize the hierarchical relationship between clusters. The height of the links between clusters represents the distance or dissimilarity between them.