Extract, Transform and Load (ETL)

Extract, Transform and Load are three typical database operations that are often combined into one tool to pull data out of one system and place it into another system with a different format or layout. Data manipulation and aggregation are important tasks in ETL, which are often necessary before a machine learning method can be applied to generate predictive or description data models.

Useful KNIME nodes for data I/O, manipulation, aggregation and visualisation

#	Type	Node	Task
1	Data Aggregation	Concatenate	Vertical merge: Concatenate two datasets (or data tables) A and B into one table C for one or more equal (matched) number of attributes of A and B.
2	Data Aggregation	Joiner	Horizontal Merge: it joins two data tables A and B into one table C for one or more matched attributes of A and B. It corresponds to the database operation "join".
3	Data Aggregation	GroupBy	Contingency table (cross tabulation): it creates a table with summary statistics by grouping data rows by the distinct combination of values of two or more selected columns and computing an aggregation function (count, average, sum, etc.) over the values of one selected column.
4	Data Aggregation	Pivoting	Contingency table (cross tabulation): see above. This node provides additional output table showing the aggregation over the rows/columns of the contingency table.
5	Data Aggregation	Crosstab	Contingency table (cross tabulation): see above. This node has an interactive view and allows more advanced statistical analysis than the previous two.
6	Data Manipulation	Category to Number	Converting String attribute to numeric
7	Data Manipulation	Cell Replacer	Replace a cell value with another value
8	Data Manipulation	Column Filter	Include/Exclude rows in the data table using a test or a regular expression. Missing values CANNOT be manipulated with this node.
9	Data Manipulation	Column Rename	Change the name of the column (column header)
10	Data Manipulation	Column Resorter	Change the order of the columns in the data table
11	Data Manipulation	Column Splitter	Partition the table column-wise. This node splits the columns of the input table into two output tables.
12	Data Manipulation	Extract Column Header	Extract the column names (headers) of a data table
13	Data Manipulation	Insert Column Header	Updates column names of a table according to the mapping in second dictionary table.
14	Data Manipulation	Missing Value	Missing values in any numeric (and/or string) column are replaced using with a specified function (min, max, average, most frequent, etc.). Numeric and string columns are managed separately.
15	Data Manipulation	Row Filter	Include/ Exclude rows in the data table based on some criterion. Missing values can be eliminated by removing the entire row.

16	Data Manipulation	Row Splitter	Partition the table row-wise. It splits the rows of the input table into two output tables. It is similar to the Row Filter node, except that it has an additional output providing the rows that are filtered out.
17	Data Manipulation	RowID	Replace the row IDs (row headers) of the input table with one of the columns in the table.
18	Data Manipulation	Sorter	This node sorts the rows according to user-defined criteria (increasing, decreasing, etc.).
19	Data Manipulation	Transpose	Transposes the entire input table by swapping rows and columns.
20	Data Manipulation	Math Formula	Apply mathematical operations to numerical values of selected columns. See the video provided in Bb.
22	Data Manipulation	String Manipulation	Apply a transformation to string values.
21	Data Manipulation	Rule Engine	Apply rule-based transformation (e.g., IF X AND Y THEN Z) to a selection of one or more columns. See the video provided in Bb.
23	IO	CSV Reader	Read CSV files from a location (relative to KNIME workspace or local computer space or from an URL)
24	IO	CSV Writer	Write a CSV file to a location (relative to KNIME workspace or in local computer space)
25	IO	File Reader	Read a file (irrespective of type) from a location (relative to KNIME workspace or local computer space or from a URL)
26	Plot	Bar Chart	Create a bar chart for a comparison of one or more columns
27	Plot	Box Plot	Create a box chart for a comparison of one or more columns
28	Plot	Histogram	Create a histogram for the comparison of one or more variables distributions (frequency) in the datasets
29	Plot	Interactive Table	Show the data table into an interactive view
30	Plot	Line Plot	Create a line chart for a comparison of one or more variables
31	Plot	Parallel Coordinates	Create a chart with parallel vertical axes, one for each attribute. A line in the plot corresponds to a record (row).
32	Plot	Pie Chart	Create a pie chart for a comparison of one or more variables distributions in the datasets
33	Plot	Scatter Matrix	Create a scatter plot to show the correlation between all combination of attribute pairs in a matrix form
34	Plot	Scatter Plot	Create a scatter plot to show the correlation between the two variables
35	Plot	Color Manager	Assign a colour to an attribute (column) to visualise different nominal or numerical values in the dataset.
36	Plot	Size Manager	Assign a marker size to an attribute (column) to visualise different nominal or numerical values in the dataset.
37	Plot	Shape Manager	Assign a marker shape to an attribute (column) to visualise different nominal (only) values in the dataset.