Cheat sheet: Data wrangling with KNIME Analytics Platform

Open for Innovation
KNIME

Access data

CSV Reader: Reads a CSV file from either your local file system or another connected file system. Click the three dots in the lower left corner to add a dynamic connection input port to connect to an external file system, like Amazon S3, Azure Blob Storage, etc.

Excel Reader: Reads sheet(s) from one or more Excel files. One sheet from each Excel file. A loop can be used to read multiple sheets from one Excel file.

Table Reader: Reads data from a .table file. The .table files are organized using a KNIME proprietary format, including the full file structure, and are optimized for space and speed - providing maximum performance with minimum configuration.

SAP Reader (Theobald Software): Loads data from various SAP systems (e.g. SAP S/4HANA, SAP BW, SAP R/3).

Amazon S3 Connector: Connects to Amazon S3 and points to a working directory (with a UNIX-like syntax, e.g., /mybucket/myfolder/myfile). Allows downstream reader nodes to access data from Amazon S3 as a file system.

Common settings of Reader and Writer nodes

File path: All Reader and Writer nodes require a file path. The file path can be expressed as an absolute path in the local file system, a relative path to a key location in the current KNIME installation, or a path defined in an external file system if such a connection is used.

Multiple files: Reader nodes can read and concatenate multiple files, according to a selected file extension or file name pattern.

Transformation tab: Reader nodes include a Transformation tab for renaming, filtering, re-ordering, and type changing of the columns.

Databases

DB Connector: Connects to any JDBC-compliant database. The JDBC driver must be added in the KNIME Preferences and then selected in the node configuration window.

H2 Connector: Connects to an H2 database. Similar dedicated connector nodes connect to other databases, such as MySQL or PostgreSQL.

DB Reader: Executes the input SQL query on the database and exports the results into a KNIME data table.

DB Table Selector: Creates a SQL query to access the database table selected in the configuration window. The table can be selected either via browsing the database metadata or via a custom SQL query.

Combine data

Concatenate: Concatenates the rows of all input tables by writing them below each other. This is especially useful for tables with shared column headers.

Joiner: Joins the columns of the two input tables based on one or multiple joining columns. Allows you to select between different joiner modes and to use multiple joining columns.

Value Lookup: Replaces the values in one column of the table at the top input port with values from a look up table provided at the bottom input port.

DB Joiner: Expands the input SQL query to include the join of two tables. It has a similar configuration window as the joiner node. No SQL coding required. There are more DB nodes, all expanding the input SQL query with additional SQL instructions. Besides the SQL Query node, no DB nodes require SQL coding.

Filter data

Row Filter: Filters rows in or out of the input table according to a filtering rule. The filtering rule can match a value in a selected column or numbers in a numerical range.

Rule-based Row Filter: Filters rows in or out according to a set of rules, defined in its configuration window. Rules are evaluated from top to bottom. Using TRUE as the antecedent applies the rule to all unmatched rows.

Reference Row Filter: Filters rows in or out from the top input table according to matching values in the selected column of the lower input table.

Column Filter: Filters columns in or out from the input table according to a filtering rule. Columns to be retained can be manually picked or selected according to their type, or based on a regex expression matching their name.

DB Row Filter: Expands the input SQL query to include the row filter criteria defined in the configuration window. Grouping of multiple conditions with an AND or OR conjunctions is also supported. No SQL coding required.

DB Query: Modifies the input SQL query using custom SQL. The input SQL query is represented by the place holder #table#.

Write data

Excel Writer: Writes the input table(s) to sheet(s) in an Excel file (XLS or XLSX). Click the three dots in the lower left corner to add a dynamic sheet input port to write multiple data tables into multiple sheets.

CSV Writer: Writes the input data table to a CSV file. Click the three dots in the lower left corner to add a dynamic connection input port to write to an external file system, like Amazon S3, Azure Blob Storage, etc.

Send to Tableau Server: Uploads the input table to a Tableau server for reporting.

Send to Power BI: Uploads the input table to Microsoft Power BI for reporting.

DB Writer: Inserts the data rows from the top input port into a table in the database specified by the input connection port. If the database table does not exist it will be created.

DB Connection Table Writer: Writes the resulting rows from the input SQL query into a new table inside the database.

Date&time

String to Date&Time: Parses the strings in the selected columns according to a date/time format and converts them into Date&Time cells. Four Date&Time forms are supported: only date, only time, date&time, and date&time plus time zone.

Date&Time-based Row Filter: Extracts rows where the time value in the selected column lies within a given time window. The time window is specified either by a start and /or an end date or by a start date and a duration.

Date&Time Difference: Calculates the difference between two date&time objects e.g., from two selected columns, from a selected column and a fixed value, from a selected column and the current execution time, or from one cell and the cell in the previous row for a selected column.

Extract Date&Time Fields: Extracts selected time and date fields from a selected column of type date&time and appends their values in new columns.

Clean data

Missing Value: Defines and applies a strategy to replace missing values in the input table - either globally on all columns, or individually for each single column.

Duplicate Row Filter: Detects duplicate rows and applies the selected operation, e.g. removes duplicate rows. Duplicates are rows that have the same value in all selected columns.

Numeric Outliers: Detects and treats numerical outliers for each of the selected columns individually using the interquartile range (IQR).

Reshape and aggregate data

GroupBy: Groups the rows of a table by the unique values in selected columns and calculates aggregation and statistical measures for the defined groups. Despite its simple name, it offers powerful functionality and has many unsuspected usages.

Pivot: Extends the aggregation functionality of the GroupBy node by creating an output table with columns and rows for the unique values in the selected input columns. The unique values of the grouping columns become rows and the unique values of the pivoting columns become columns.

Category to Number: Maps the categorical values in the selected columns to integer values and exports the mapping rules to the model output port. The Category to Number (Apply) and Number to Category (Apply) nodes apply the mapping rule in both directions.

One to Many: Creates one new column for each value in the selected input column. These values become the column headers. Cells in the newly created columns are set to 0 if the value is not present, otherwise 1. This type of encoding is called one-hot vector.

Table Transposer: Converts the rows to columns and the columns to rows.

Table Manipulator: Performs several transformations at once, such as renaming, filtering, re-ordering and type changing, on the input columns. By adding dynamic ports it can replace a concatenate node.

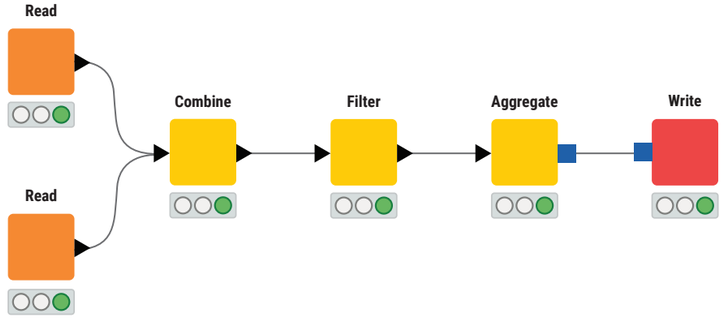
Cell Splitter: Splits values in the selected column into two or more substrings, as defined by a delimiter match. A delimiter is a defined character, such as a comma, space, or any other character or character sequence.

Ungroup: Ungroups a collection-type cell by creating one row for each value in the collection cell. Other columns from the input table are left unaltered.

Unpivot: Stacks the cells of the selected value columns into one column. The cells of the selected retained columns are appended to the corresponding output rows.

Sorter: Sorts the table in ascending or descending order based on the values of one or more columns.

Value Counter: Counts the number of occurrences of all values in a selected column from the input table.



Data types & conversions

- **String**: Sequence of characters, e.g. "This is a string"
- **Integer**: Whole real valued number, e.g. -100 or 345
- **Double**: Real valued number, e.g. -0.432 or 45.39
- **Date&Time**: A data format for date, time, date&time, or date&time plus time zone.
- **Boolean**: Two possible values only, e.g. TRUE and FALSE
- **Collection Cell**: Collection of multiple values of either the same or different types e.g., can be a list of values or a set of values. In a set each value occurs only once.
- **Document/Image**: KNIME Analytics Platform supports many more data types like text documents, images, fingerprints, etc.

Number to String: Converts the data type of the selected columns from a number format, e.g. integer or double, to string.

String to Number: Converts the data type of the selected columns from string to either double or integer.

Create columns

Math Formula: Implements a number of math operations across multiple input columns. The math operations can be applied to multiple columns with the Math Formula (Multi Column) node.

Rule Engine: Applies a set of rules to each row of the input table. Rules are applied from top to bottom. The first rule that matches is used.

Counter Generation: Creates a new column with a counter. The start value and step size are defined in the configuration window.

String Manipulation: Performs operations on string values in columns, such as combining two or more strings together, extracting one or more substrings, trimming blank spaces, and so on.

String Replacer: Replaces values in a selected string column if they match a defined pattern.

Column Expressions: Combines the functionality of the Math Formula, Rule Engine, and String Manipulation nodes. More than one expression can be defined to modify or add multiple columns at the same time.

Dynamic port

Dynamic ports: Additional input ports can be added by clicking the plus on left side of a node.

Format excel sheets

The Continental Nodes for KNIME extension allows you to automatically format an existing Excel sheet. The key is an additional data table of the same size as the original Excel sheet, where each cell contains one or more comma separated tag values e.g., header, border, etc. Based on these tags, the XLS Formatter nodes add new formatting instructions to the existing instructions, as available at the lower (optional) input port.

XLS Control Table Generator: Transforms the input table to an XLS Control Table, meaning it exchanges the column names to A, B, C, ... and the row IDs to 1, 2, 3, ... It is the kickoff node to collect formatting instructions for an Excel sheet and feeds all XLS formatter nodes.

XLS Background Colorizer: Adds background color and/or pattern fill formatting instructions to all cells with a specified tag in the XLS Control Table at the top input port.

XLS Border Formatter: Adds border formatting instructions for a given range specified by a tag in the XLS control table at the top input port.

XLS Cell Merger: Adds formatting instructions to merge all cells with a specified tag in the XLS control table at the top input port.

XLS Conditional Formatter: Adds formatting instructions to color cell backgrounds according to their numeric value for all cells specified by a tag in the XLS control table at the top.

XLS Formatter (apply): Applies all formatting instructions to an existing Excel sheet.

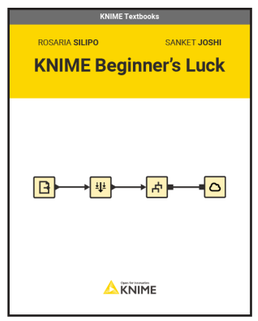
Resources

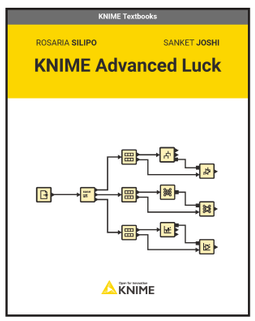
- **KNIME Press:** Access various data science books and other cheat sheets at ***knime.com/knimepress***, including beginner and advanced topics.
- **KNIME blog**: Engaging topics, challenges, industry news, & knowledge nuggets at ***knime.com/blog***.
- **Self-paced courses:** Take our free online self-paced courses to learn about data analysis, data engineering, or data science with KNIME (with hands-on exercises) at ***knime.com/learning***.
- **KNIME Community Hub:** Browse and share workflows, nodes, and components or access collection pages for dedicated topics at ***hub.knime.com***.
- **KNIME Forum**: Join our global community & engage in conversations at ***forum.knime.com***.
- **KNIME Business Hub:** For team-based collaboration, automation, management, & deployment check out KNIME Business Hub at ***knime.com/knime-business-hub***.

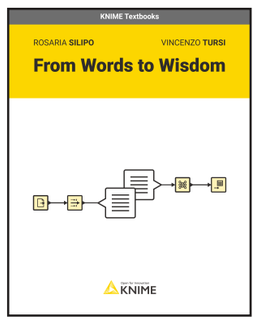
KNIME Press

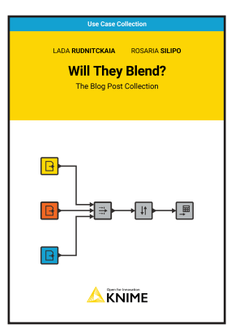
Open for Innovation
KNIME

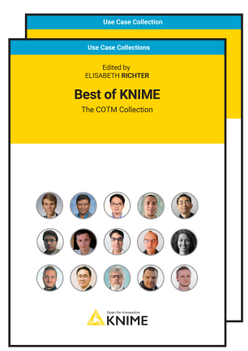
Extend your KNIME knowledge with our collection of books from KNIME Press. For beginner and advanced users, through to those interested in specialty topics such as topic detection, data blending, and classic solutions to common use cases using KNIME Analytics Platform - there's something for everyone. Available for download at www.knime.com/knimepress.

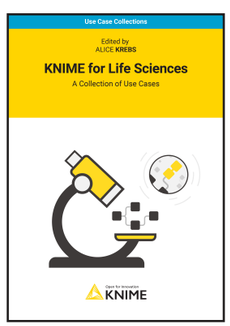


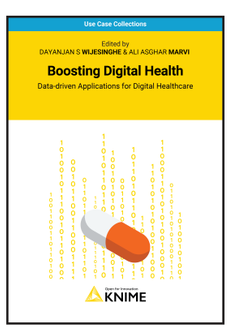


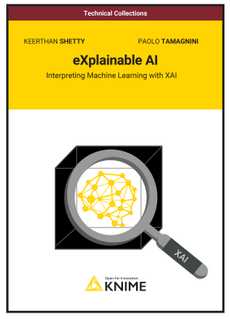


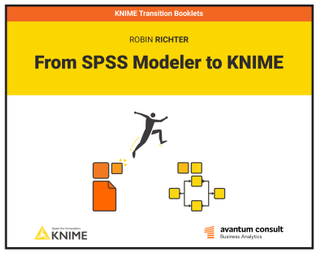


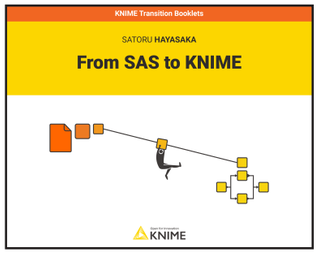


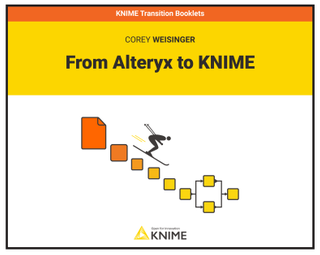












Need help?
Contact us!