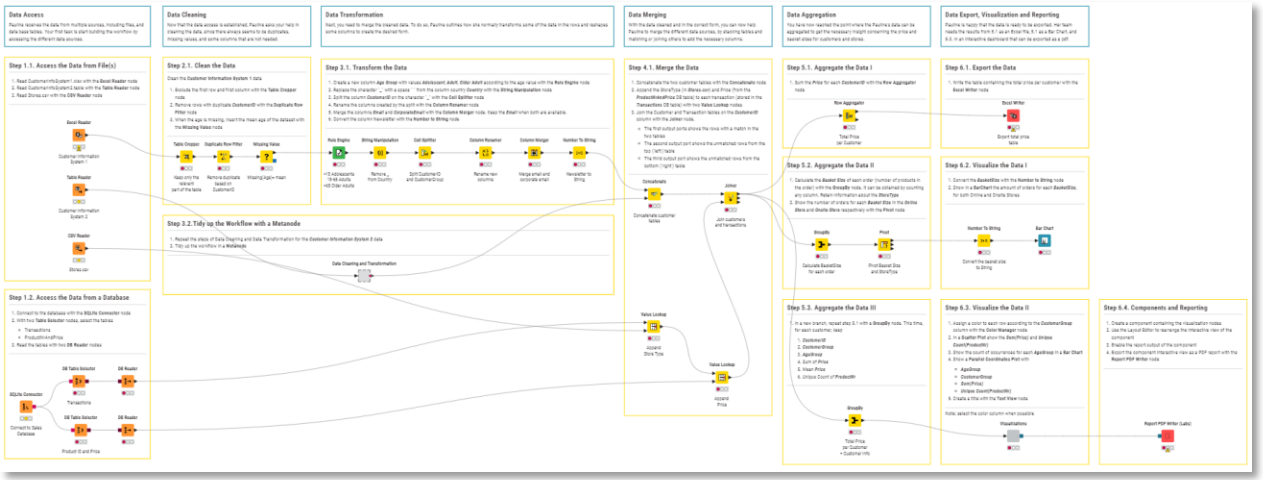
Introdução à Plataforma KNIME
Data Access

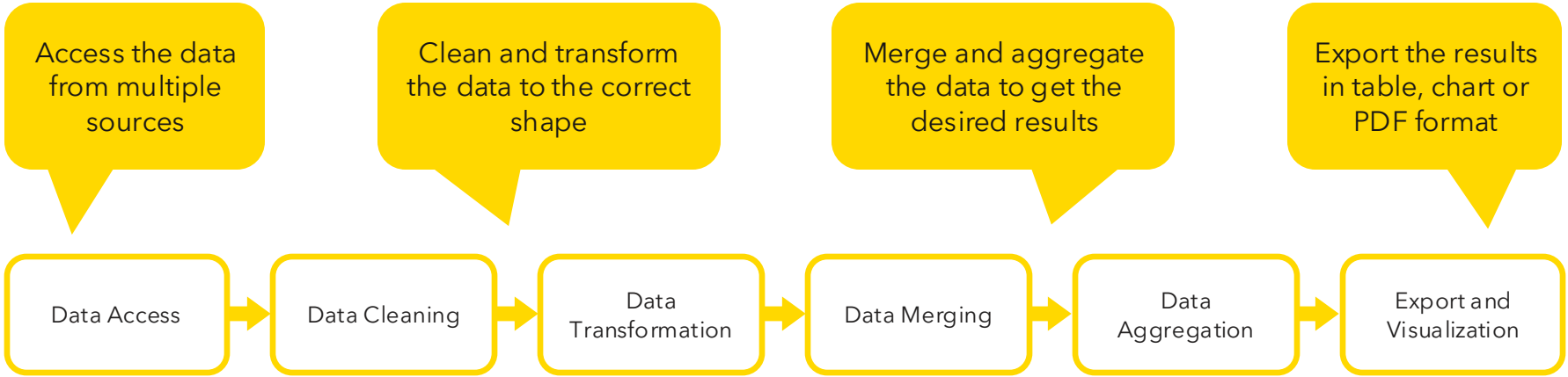
Use Case: Customer Transactions Analysis

Use Case

Pauline, do setor de Vendas, gostaria de escrever um relatório e visualizar os insights dos dados de transações de clientes que sua equipe coleta todos os meses.
Ela entrou em contato com sua equipe de dados para pedir ajuda na automatização desse processo e, assim, economizar tempo.



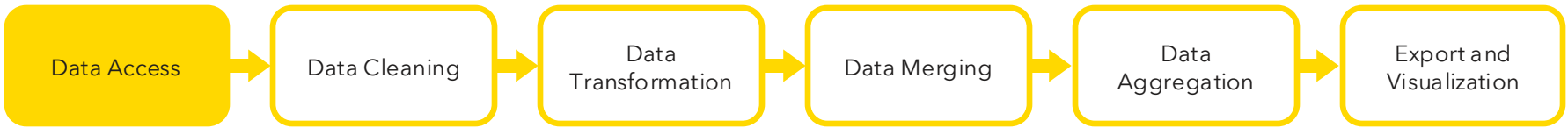
Use Case: Customer Transactions Analysis



Data Access

- The beginning of every data process
- Data can be stored in many ways
  - Locally
  - In different data format (.csv, .xls…)
  - On the cloud
  - In a database
  - …
- We need a way to access all of them

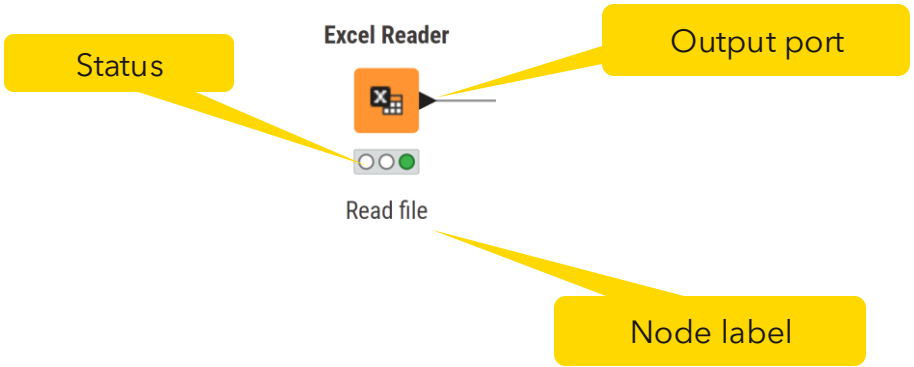
Caso de Uso
Pauline recebe os dados de várias fontes, incluindo arquivos e tabelas de banco de dados.
Sua primeira tarefa é começar a construir o fluxo de trabalho acessando essas diferentes fontes de dados.



Data Source Nodes

Typically characterized by:

- Orange color
- By default no input ports, 1-2 output ports
- Many nodes for many data formats
- Support reading from different File Systems



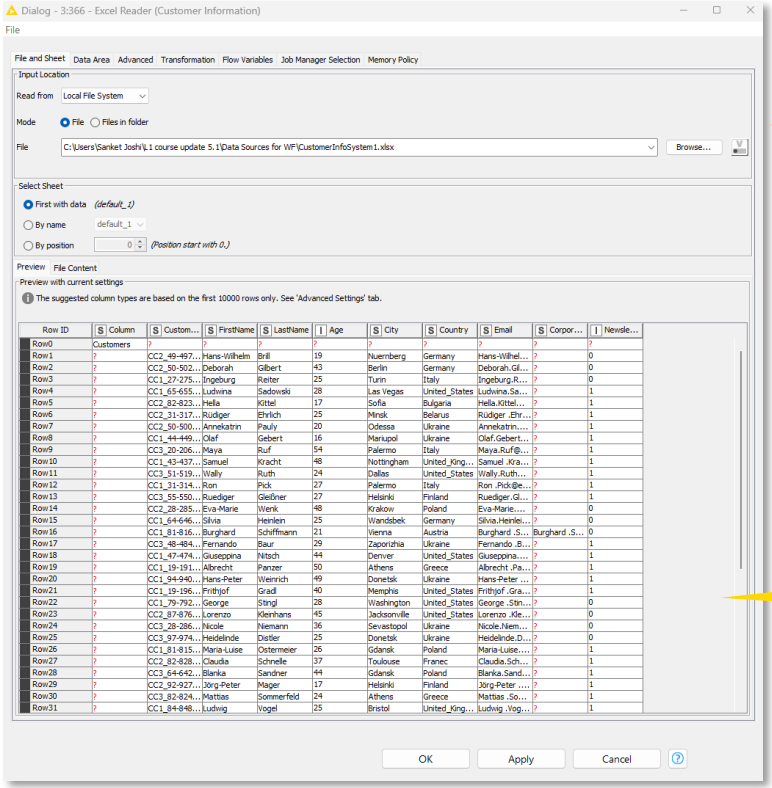
Excel Reader

- Reads .xls and .xlsx file from Microsoft Excel
- Supports reading from multiple sheets

Excel Reader

File system

Sheet specific settings

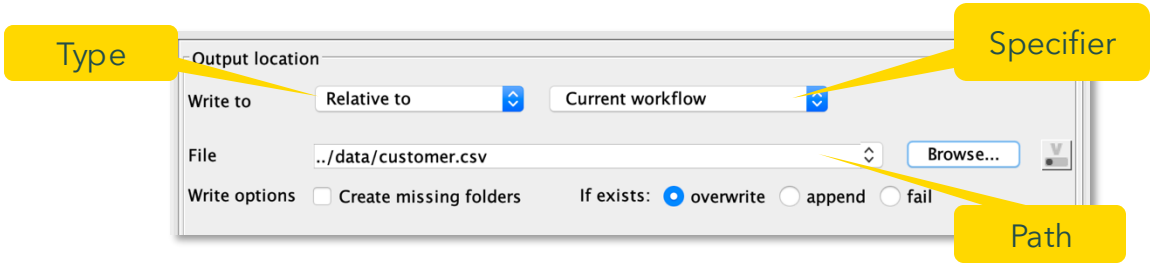


File path

Preview

Common Settings: File Path

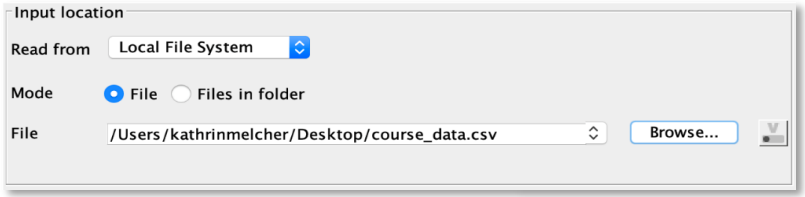
- A path consists of three parts:
  - **Type**: Specifies the file system type - e.g., local, relative, mountpoint, custom URL or connected
  - **Specifier**: Optional string with additional file system specific information - e.g. relative to which location (knime.workflow, LOCAL mountpoint…)
  - **Path**: Specifies the location within the file system



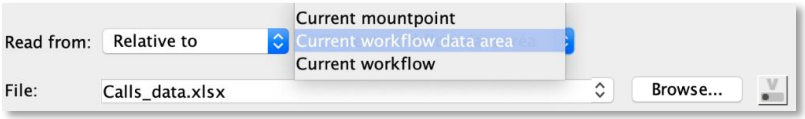
- Examples:
  - (LOCAL, , C:\Users\username\Desktop)
  - (RELATIVE, knime.workflow, file1.csv)
  - (MOUNTPOINT, MOUNTPOINT_NAME, /path/to/file1.csv)
  - (CONNECTED, amazon-s3:eu-west-1, /mybucket/file1.csv)

Common Settings: Four Default File Systems

- Local File System



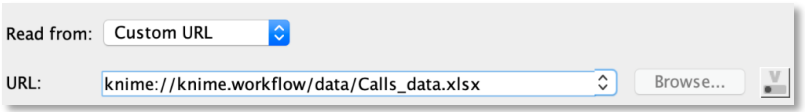
- Relative to …



- Mountpoint

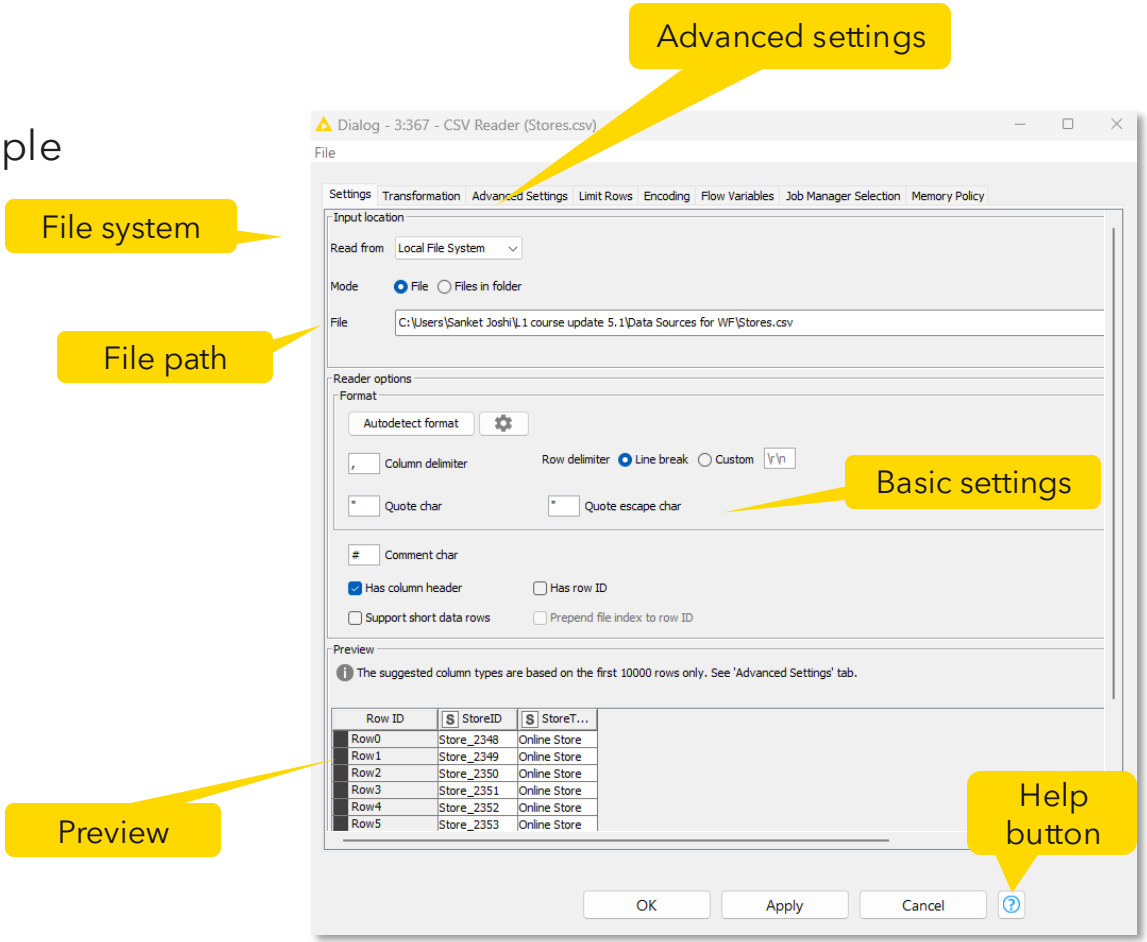
Caminho base virtual (como knime://knime.mountpoint/)

- Custom URL

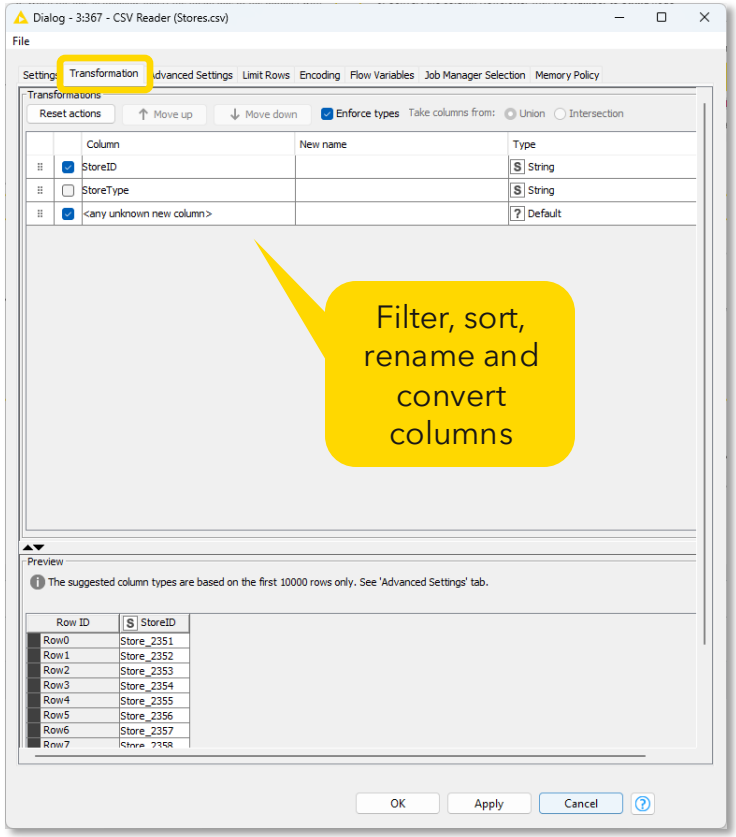


CSV Reader

- Reads either one or multiple .csv and .txt files
- Further tabs to
  - Select columns
  - Limit the rows
  - Handle quotes
  - Select encoding



CSV Reader



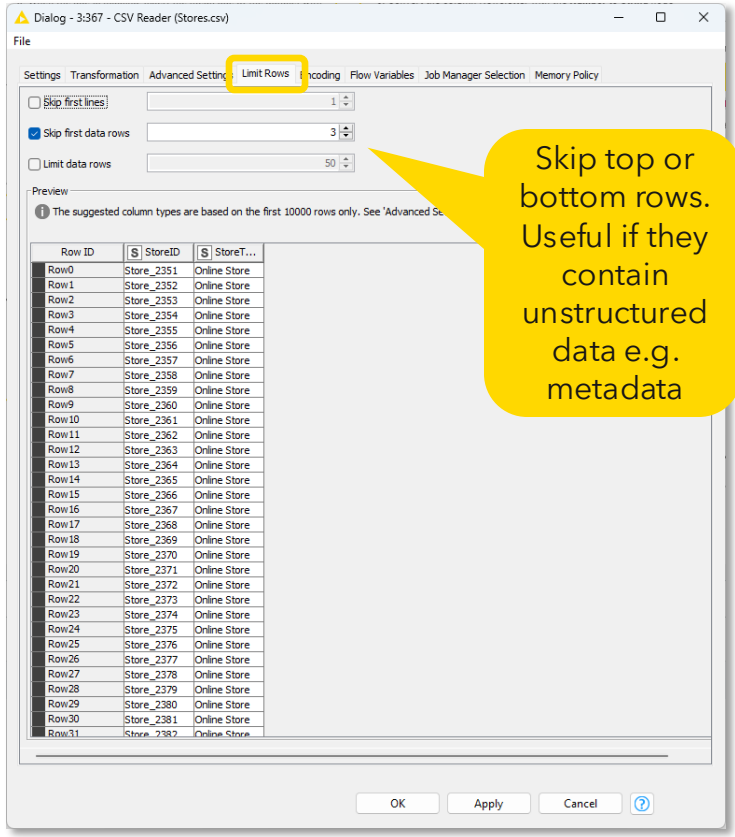


Table Reader

- Reads tables from the native KNIME Format
- Maximum performance, minimum configuration

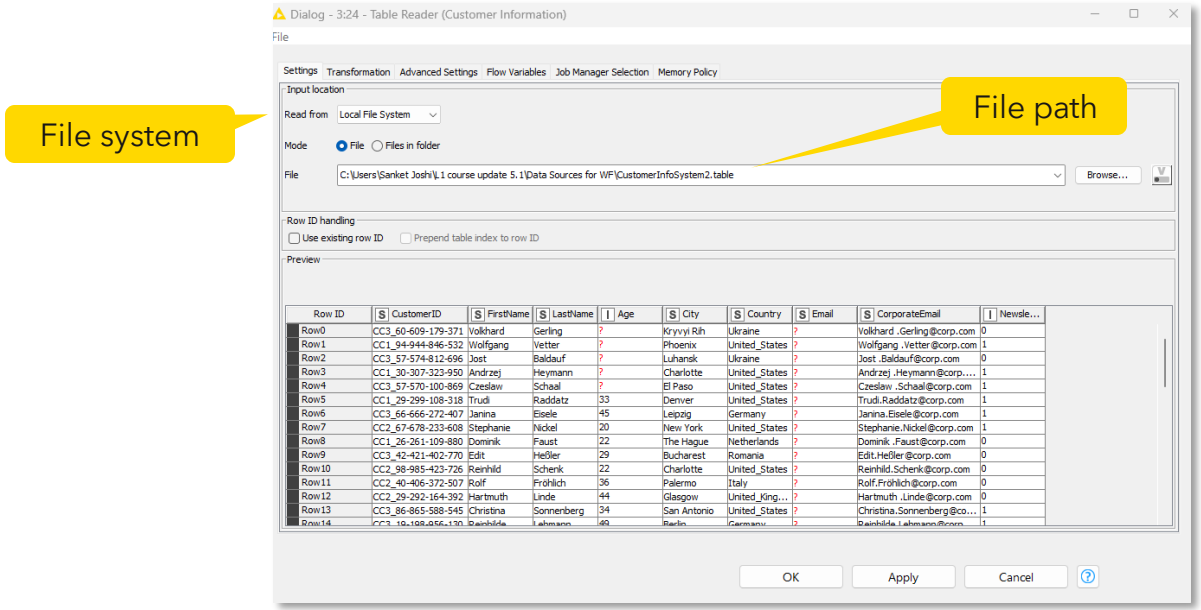
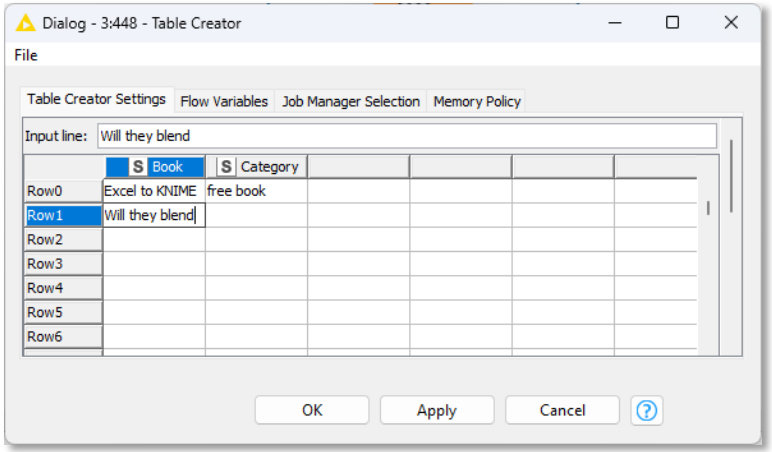


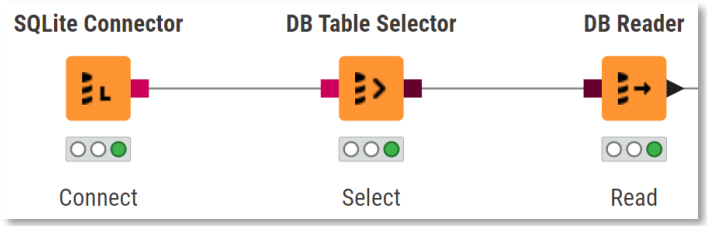
Table Creator

- Create data tables manually
- Enter data in a spreadsheet-like configuration window



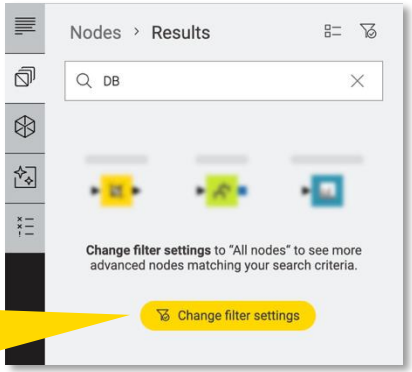
Database Connectivity

- Read data from any JDBC enabled database
- Write your own SQL or define query using dedicated nodes

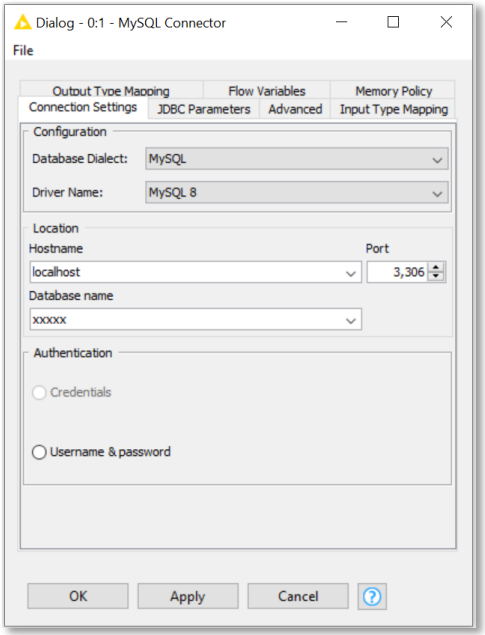


Database Connectors

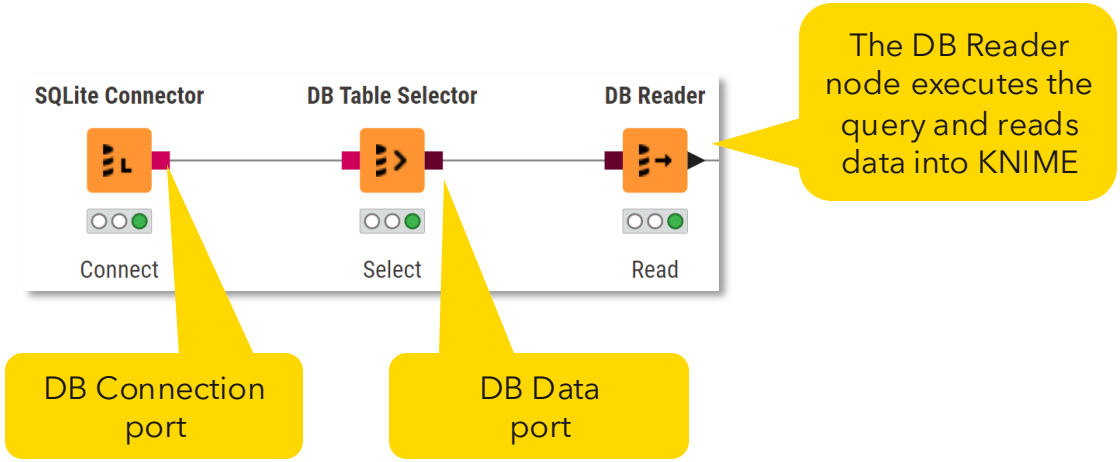
- Native: PostgreSQL, MySQL, MS SQL Server, SQLite
- DB Connector (e.g., DB2, HANA)
- Big Data: HIVE and Impala



Change the filter settings to see these nodes

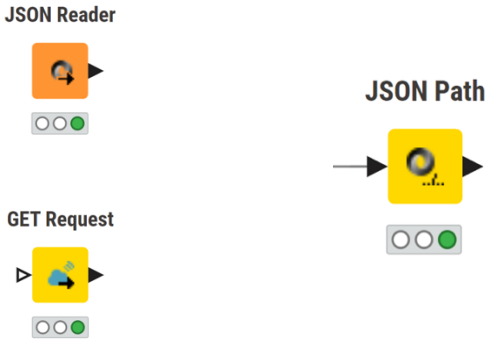


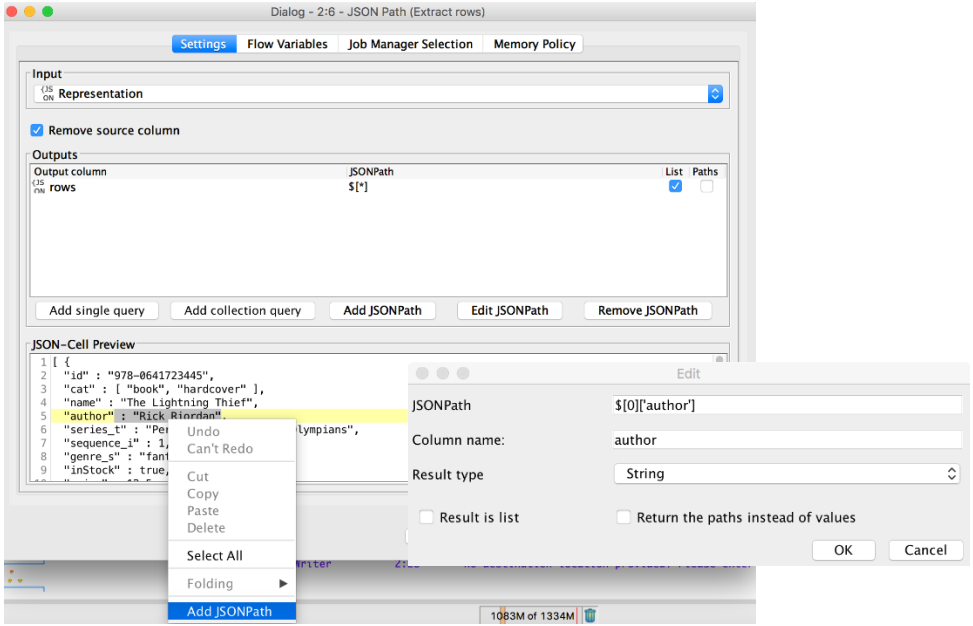
Database Query



Read JSON format

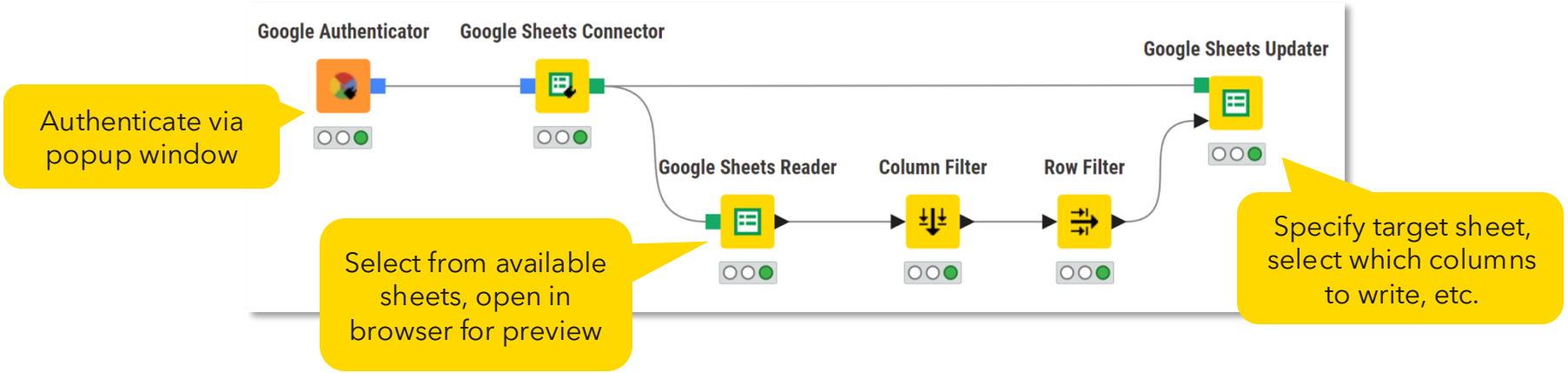
- Use the JSON Reader (or GET Request) node to get a JSON cell
- Use the JSON Path node to query the JSON file and extract parameters
  - Editor window simplifies construction of JSON queries by auto-generating them





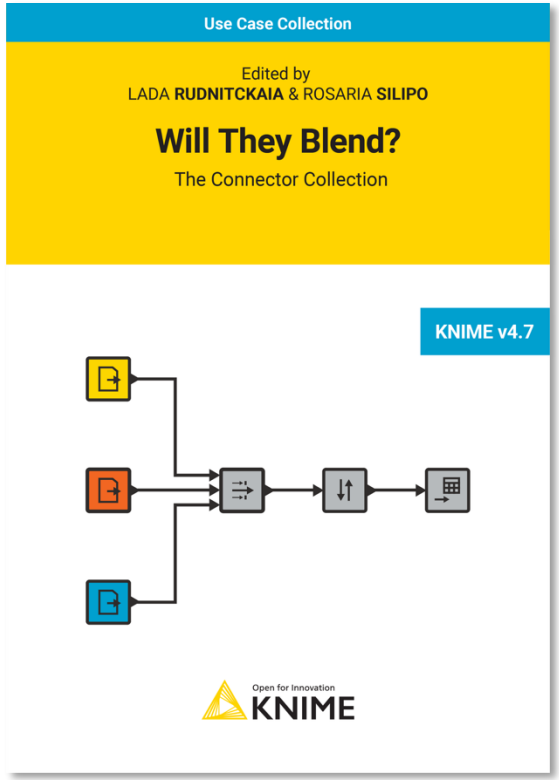
Google Sheets

- Select from available sheets on Google Drive
- Transform data in KNIME, or enrich with new data
- Create new sheet or update existing sheets
  - Allows to read from / write to specific range of sheets (e.g. A1:G10)



Other Useful Data Sources

- KNIME Analytics Platform provides many more options to access data:
  - Azure Data Lake Storage
  - Snowflake Connector
  - SMB Connector (e.g. Samba and Windows Server)
  - Python/R Source nodes
  - Tika Parser – extracts textual data from 200+ file types



Download the free book

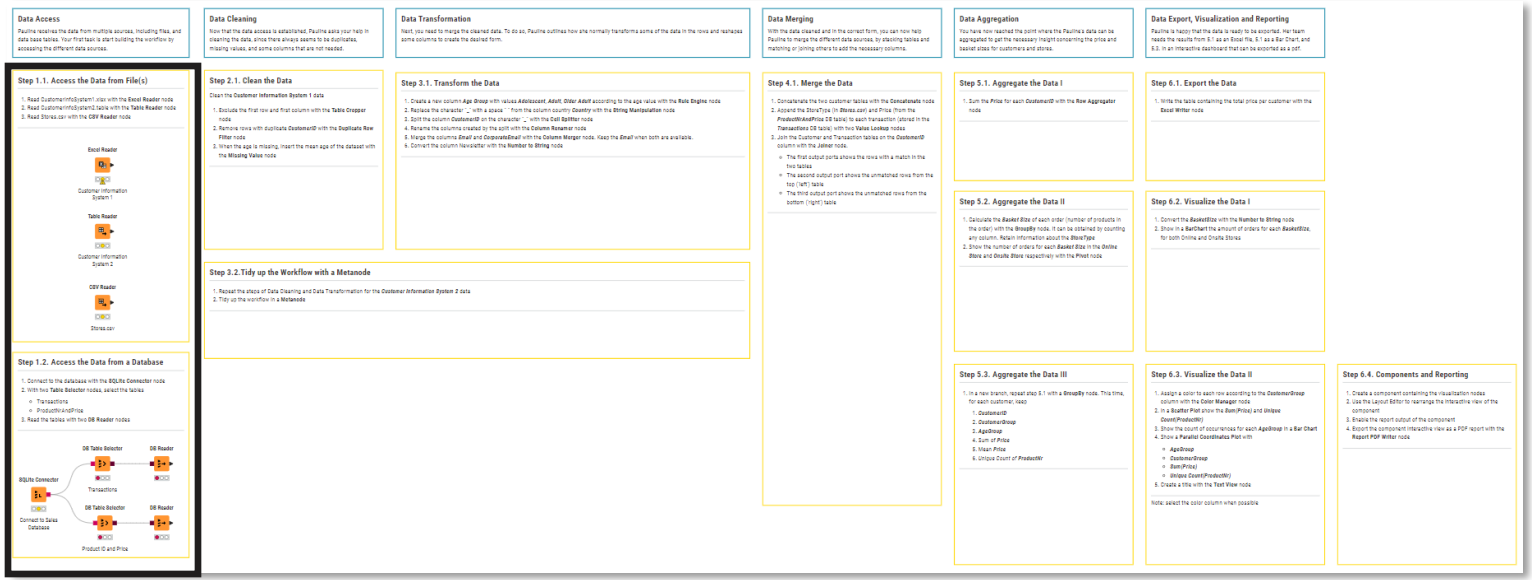
Exercises Session 1

- **Step 00**
  - Download and install KNIME Analytics Platform
  - Get familiar with the User Interface and create your first workflow
  - Download and import the training workflows for this course
  - Open the ***Customer Transactions Analysis - Exercise*** workflow and complete today's steps

You can download the training workflows from the KNIME Community Hub
hub.knime.com/knime/spaces/Education/latest/Courses

Exercises Session 1

- **Step 1.1** – Read customers and stores data from .xlsx, .table, and .csv files
- **Step 1.2** – Read transactions and products data from database



- Compare your results with the *Customer Transactions Analysis - Solution* workflow

Session 2
Data Cleaning, Data Transformation, and Workflow Documentation

Learning Objectives

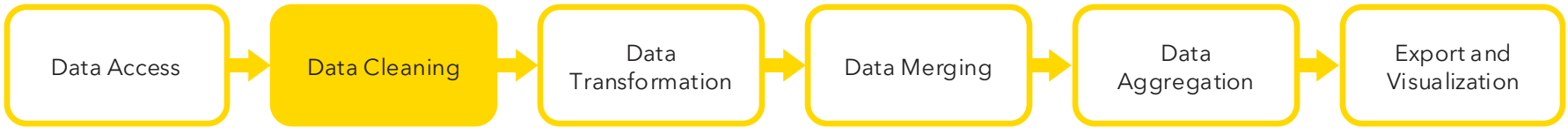
1. Filter rows and columns
2. Transform values in cells
3. Transform tables
4. List the best practices to organize and document workflows

Data Cleaning

Data Cleaning

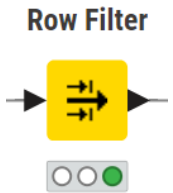
- Data are rarely clean
- Remove not useful data
- Remove repeated data
- Handle missing values

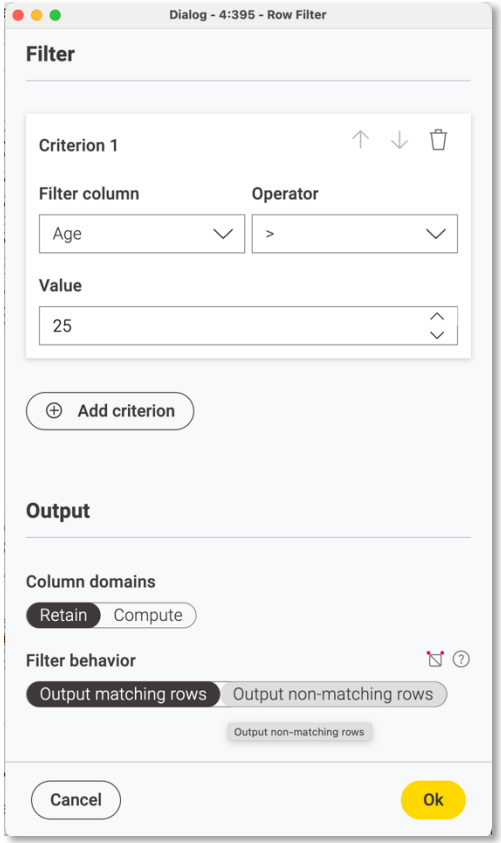
Caso de Uso
Agora que o acesso aos dados foi estabelecido, Pauline solicita sua ajuda para limpar os dados, pois há duplicatas, valores ausentes e algumas colunas que não são necessárias.



Row Filter

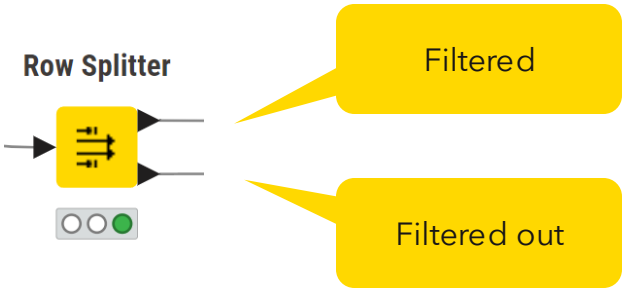
- Row filtering with include and exclude options according to certain criteria
  - Certain values or patterns in the selected column
  - Row number
  - Row ID

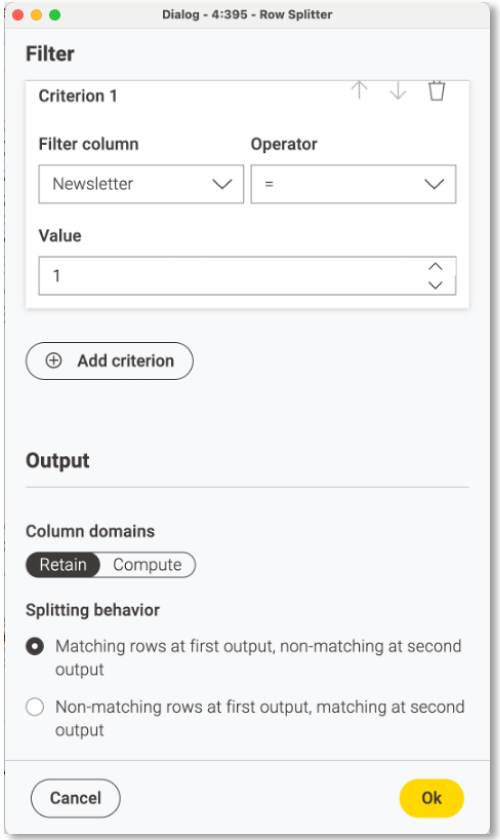




Row Splitter

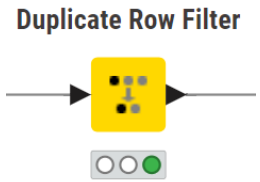
- Same configuration as Row Filter
- Filtered out rows are available in second output port

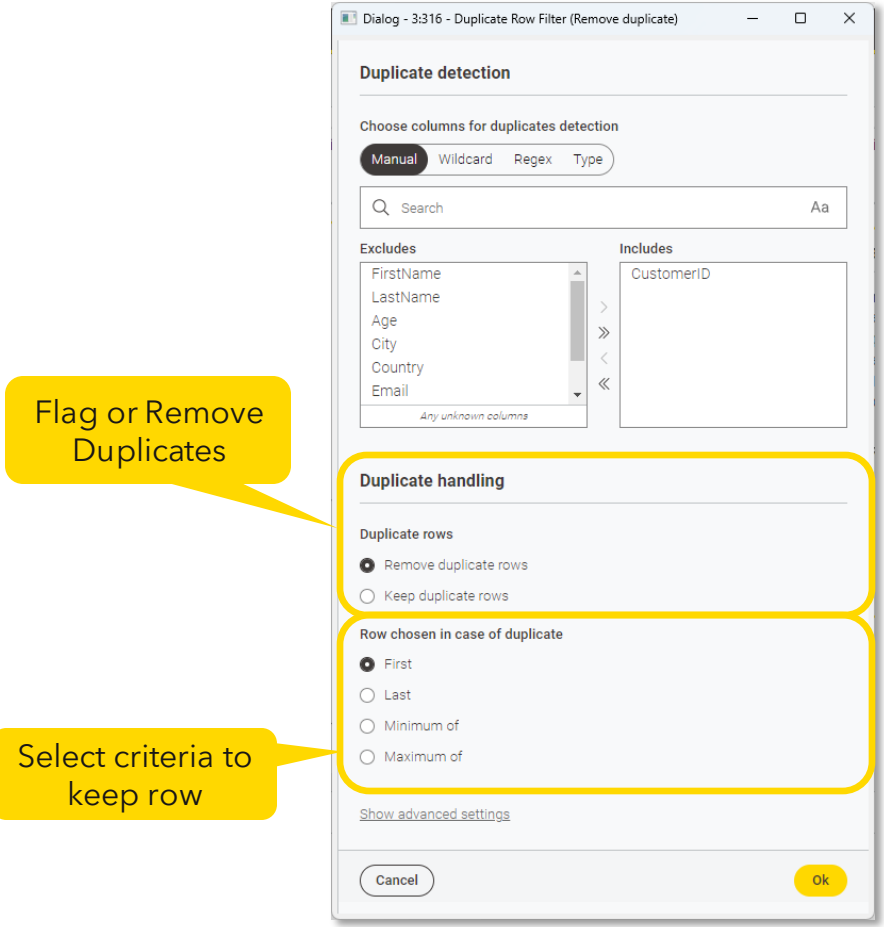




Duplicate Row Filter

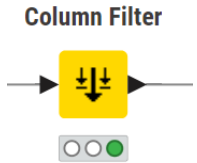
- Detect duplicate rows and apply a selected treatment
  - Select columns to check for duplicates
  - Provide options for treating duplicated values





Column Filter

- Remove columns from table



CustomerID	FirstName	LastName	Age	Newsletter
?	?	?	?	?
6589	Peter	Parker	31	yes
6768	Bruce	Banner	32	no
6925	Natasha	Romanoff	34	no

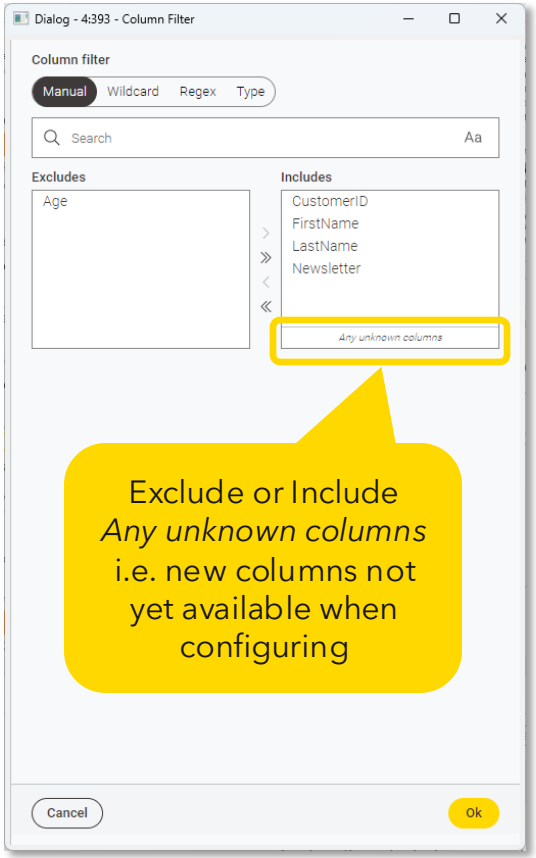
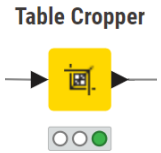
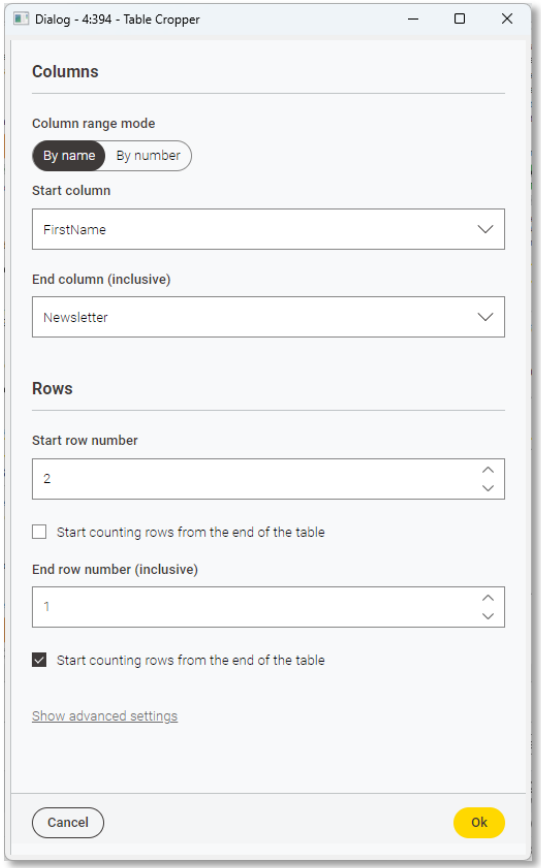


Table Cropper (recortador)

- Retain a selected, contiguous portion of a table
  - Select starting and ending rows and columns

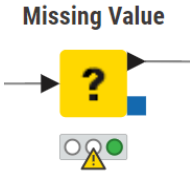


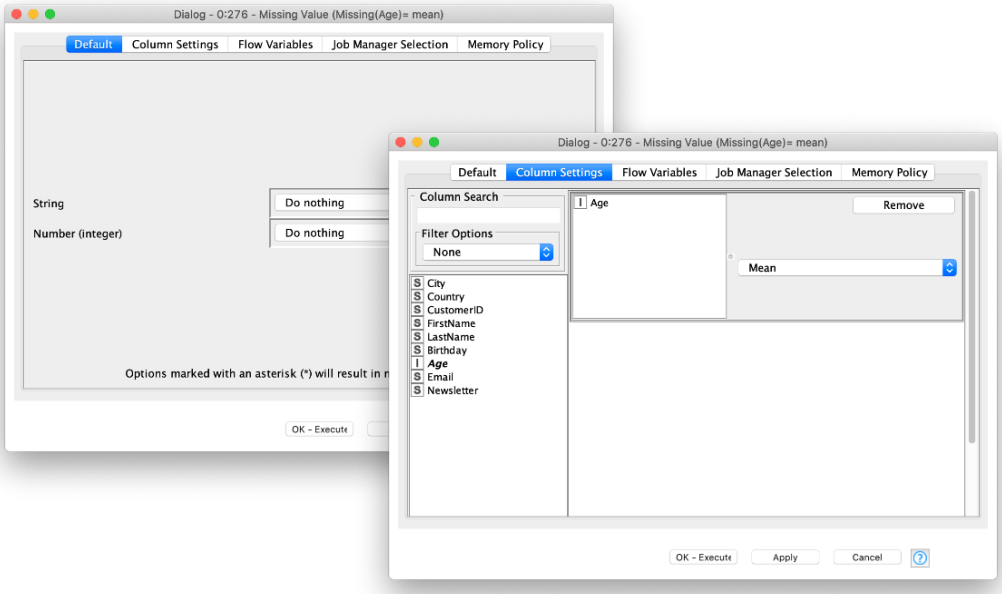
CustomerID	FirstName	LastName	Age	Newsletter
?	?	?	?	?
6589	Peter	Parker	31	yes
6768	Bruce	Banner	32	no
6925	Natasha	Romanoff	34	no



Missing Value

- Define how to handle missing values for all columns of a given type
  - Affect all columns that are not explicitly mentioned in the second tab
- Define how to handle missing values for each available column



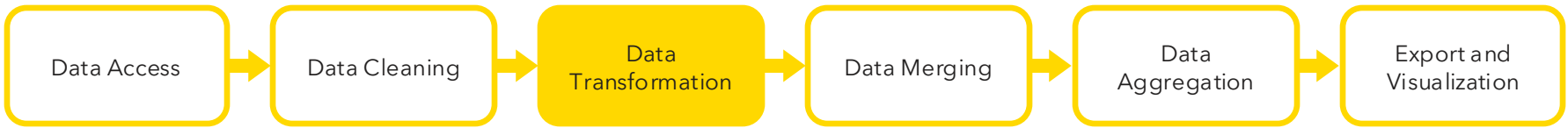


Data Transformation

Data Transformation

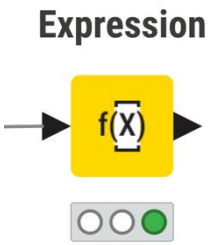
- The core of the data pipeline
- Extract more information
  - Define rules and mathematical operations
  - Transform at cell or row level
- Transform the data to the desired shape
  - Rename and resort table columns
  - Split and merge columns
  - Convert data types

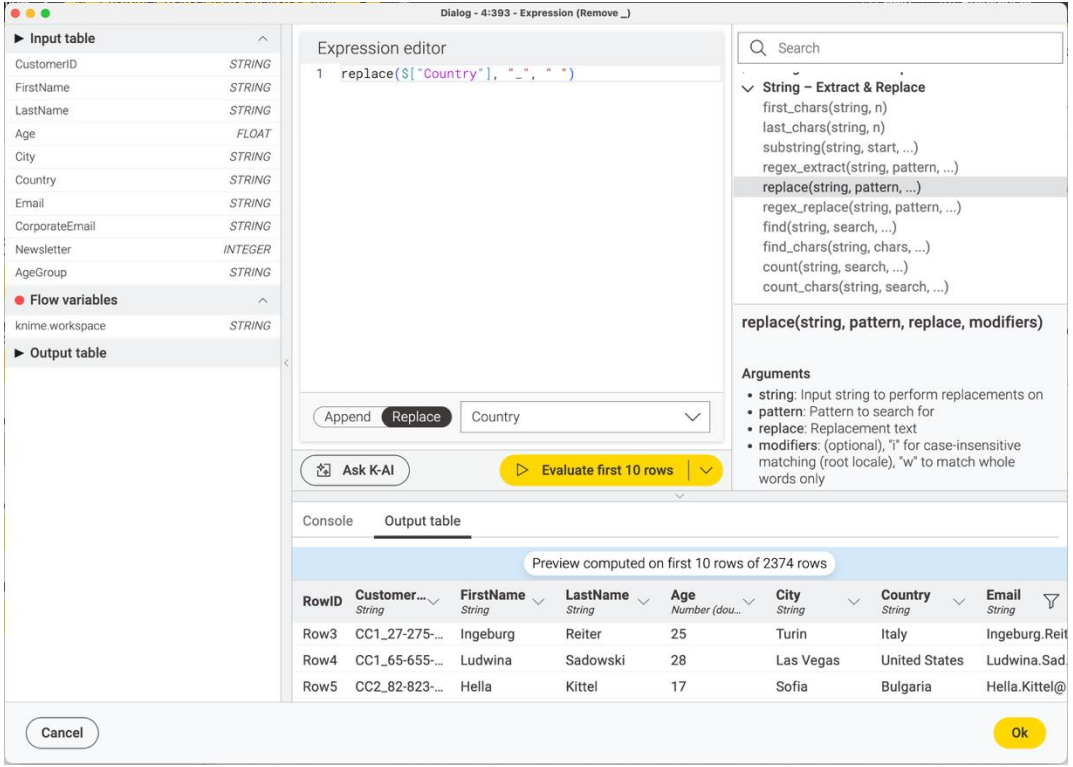
Caso de Uso
Em seguida, você precisa mesclar os dados limpos. Para isso, Pauline descreve como normalmente transforma alguns dos dados nas linhas e reorganiza algumas colunas para criar o formato desejado.



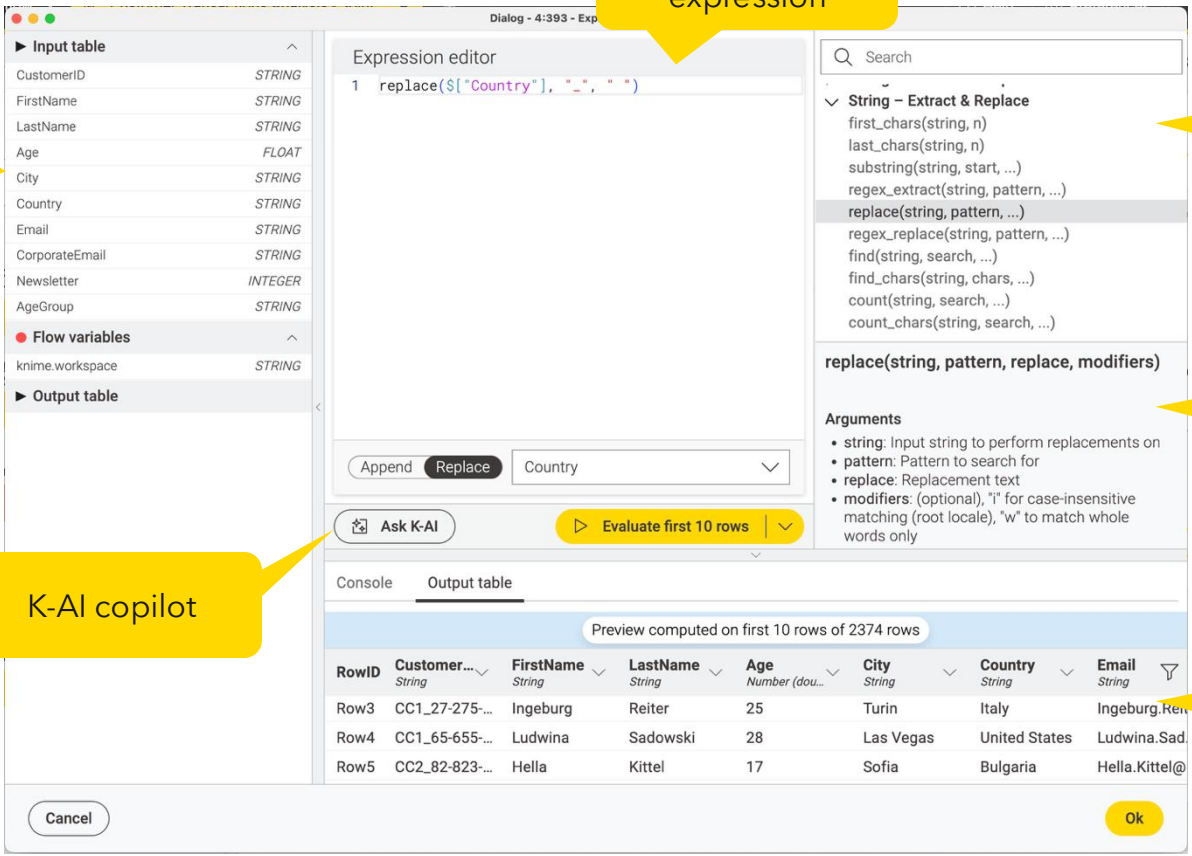
Expression

- String operations
  - Capitalize
  - Join strings
  - …
- Math operations
  - Round
  - Sum
  - Max
  - …
- Custom logic with if-clause

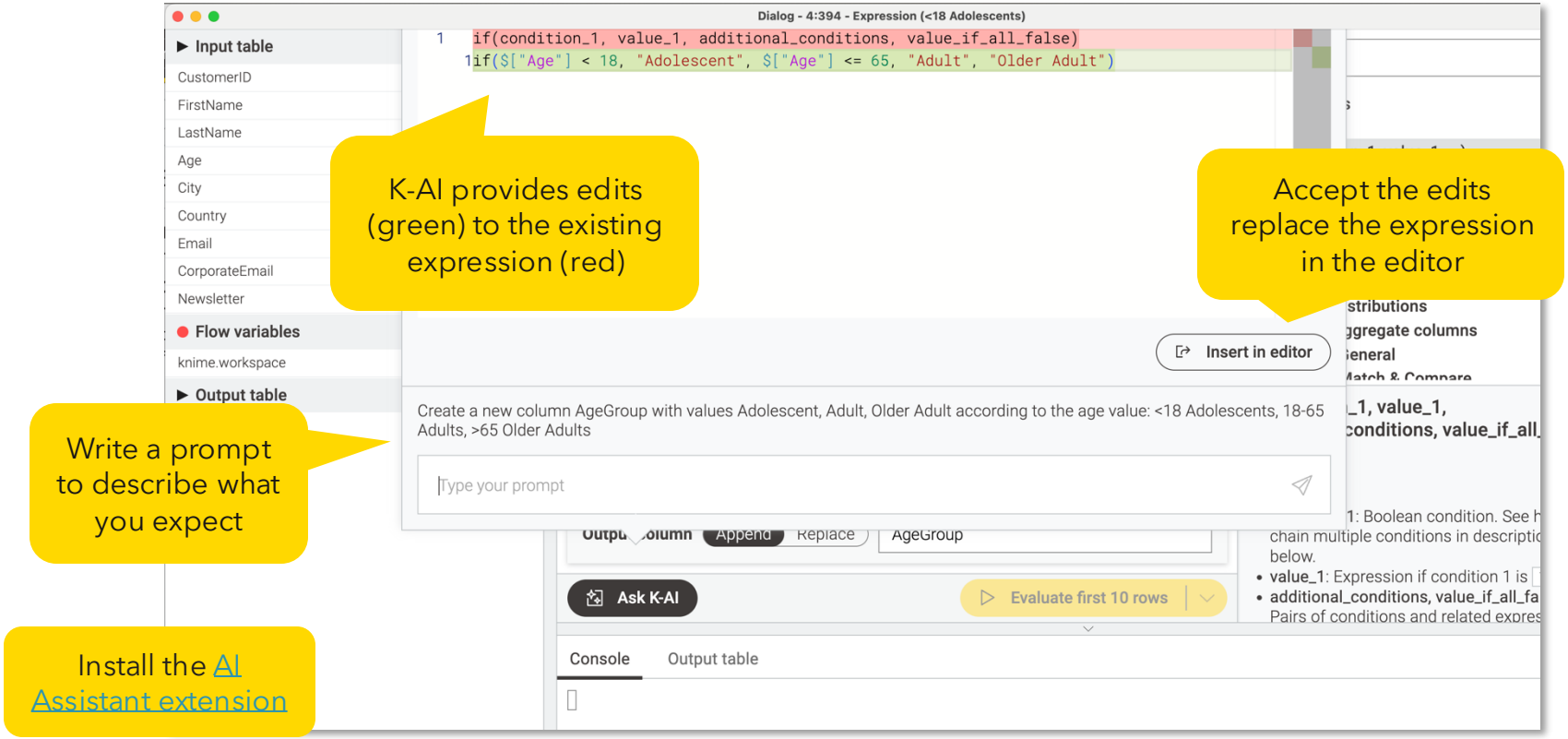




Expression

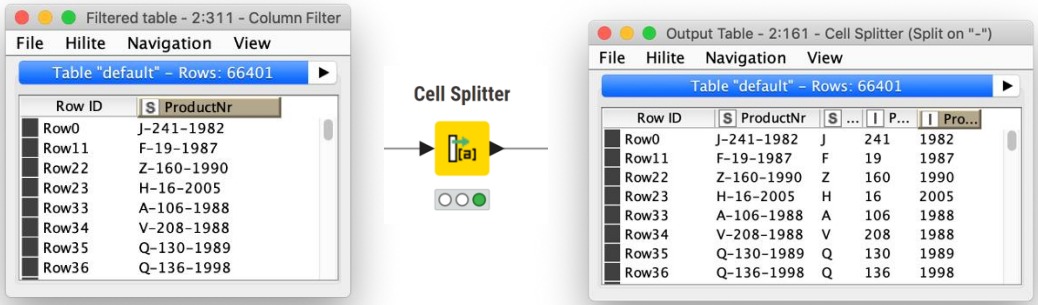


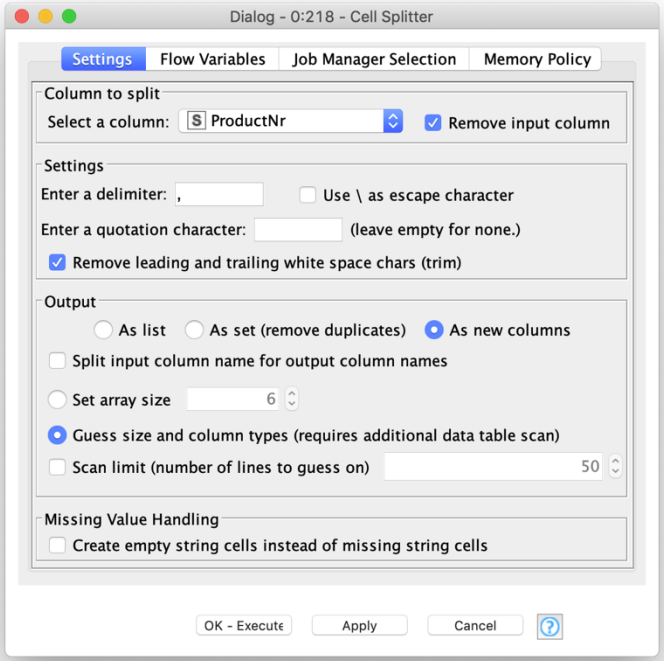
Expression with K-AI Copilot



Cell Splitter

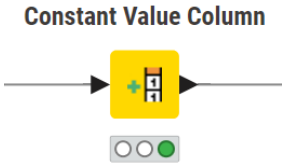
- Split the content of one column into many columns based on a delimiter

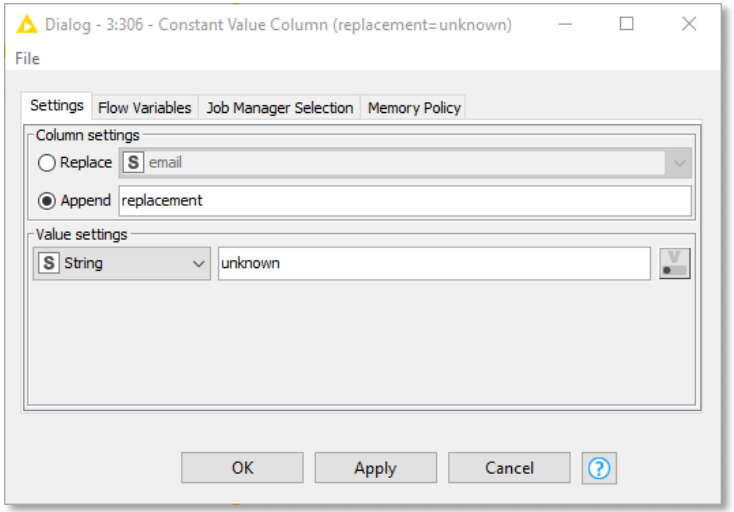




Constant Value Column

- Add or replace a column with a single constant value
- Can be used to add an empty column

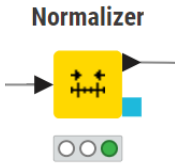




Normalizer

- Normalize values of numeric columns
  - Min-Max Normalization
  - Z-Score (Gaussian) Normalization
  - Decimal Scaling

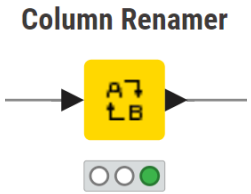
User	Age	Income
user1	34	30000
user2	20	45000
user3	59	20000
user4	60	100000

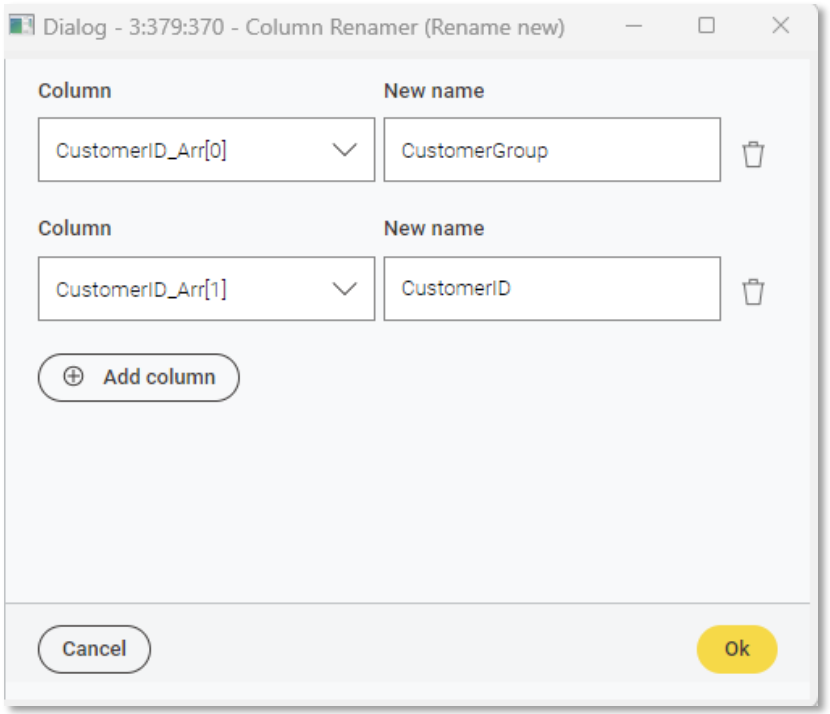


User	Age	Income
user1	0.35	0.125
user2	0	0.312
user3	0.975	0
user4	1	1

Column Renamer

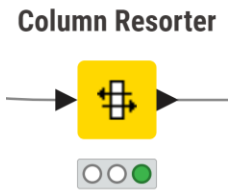
- Change name of one or more columns

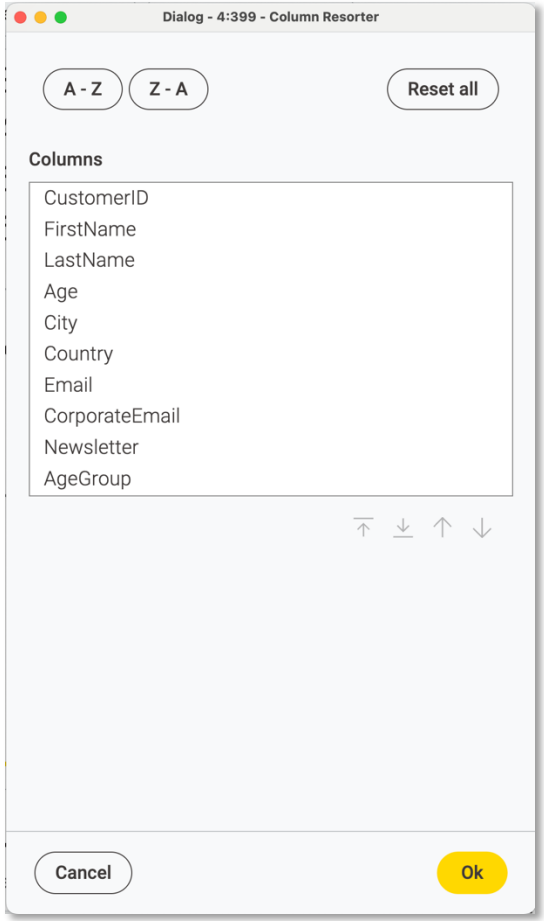




Column Resorter

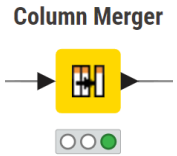
- Change the order of the columns in a table

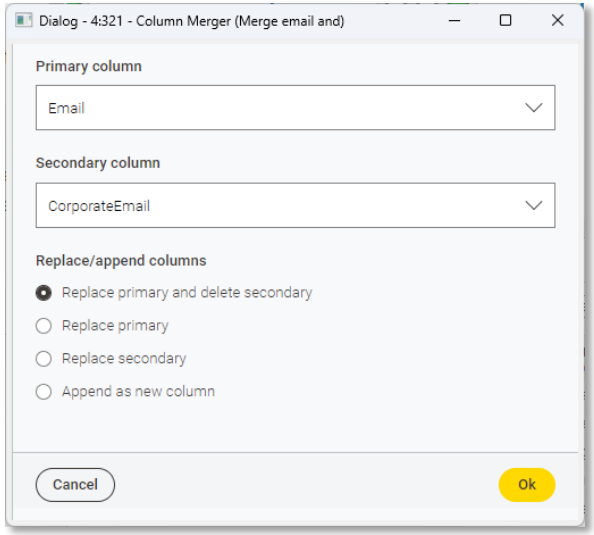




Column Merger

- Merge two columns into one by choosing the cell that is not missing



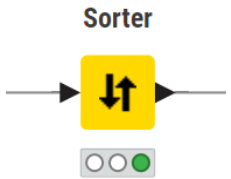


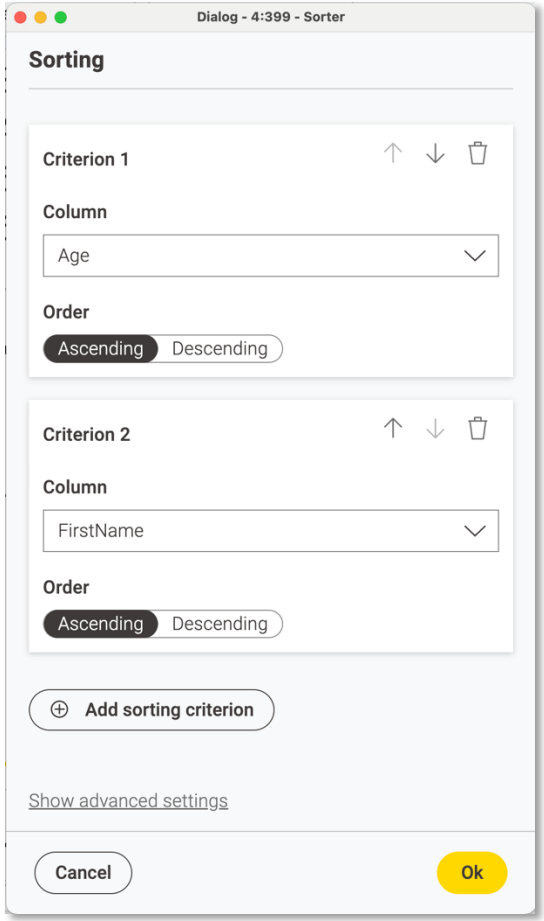
Email	Corporate Email
peter.parker@gmail.com	?
?	blackw@marvel.com
bbanner@gmail.com	thehulk@marvel.com

Email
peter.parker@gmail.com
blackw@marvel.com
bbanner@gmail.com

Sorter

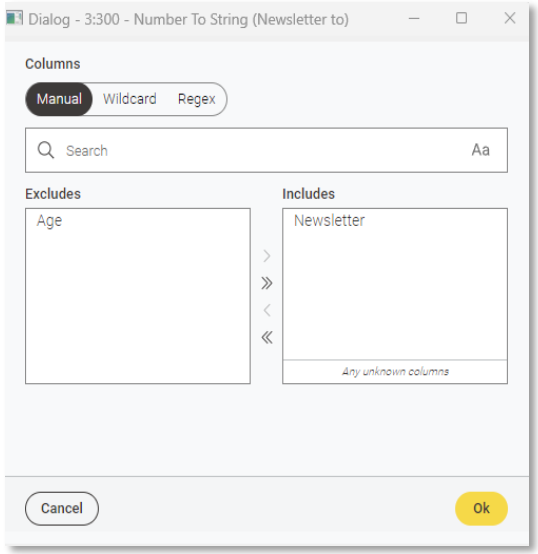
- Sort the rows based on the values of the selected column(s), either
  - ascending or
  - descending



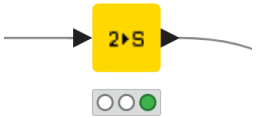


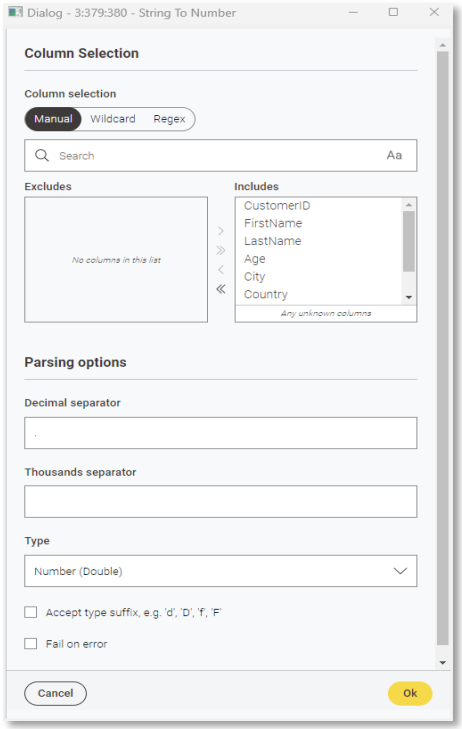
Type Conversion

- Change the data type of the selected columns



Number To String





String To Number