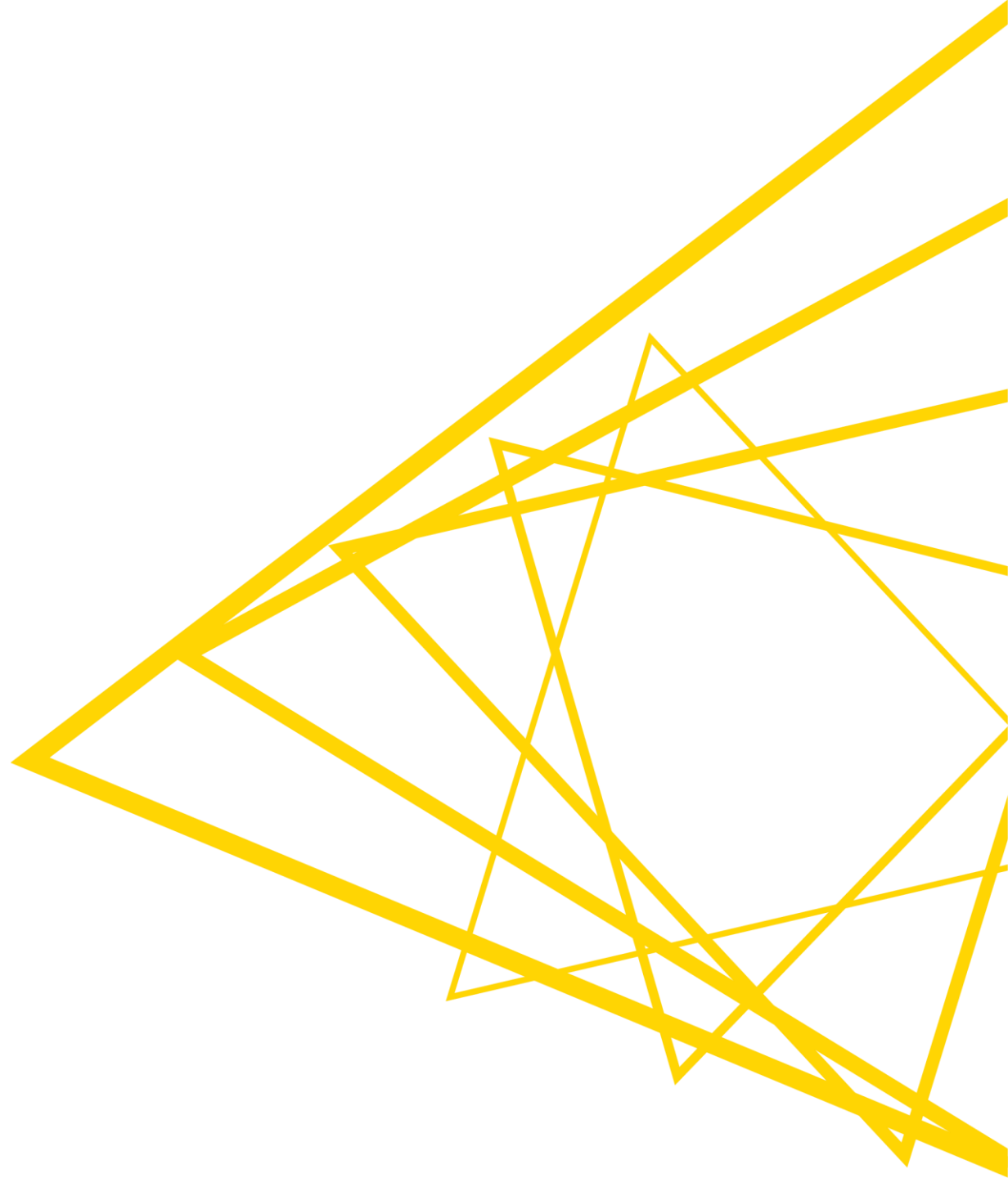




KNIME Big Data Training

education@knime.com



Overview

KNIME Analytics Platform

What is KNIME Analytics Platform?

- A tool for data analysis, manipulation, visualization, and reporting
- Based on the graphical programming paradigm
- Provides a diverse array of extensions:
 - Text Mining
 - Network Mining
 - Cheminformatics
 - Weka machine learning
 - Many integrations, such as Java, R, Python, etc.

Additional Resources

KNIME pages (www.knime.org)

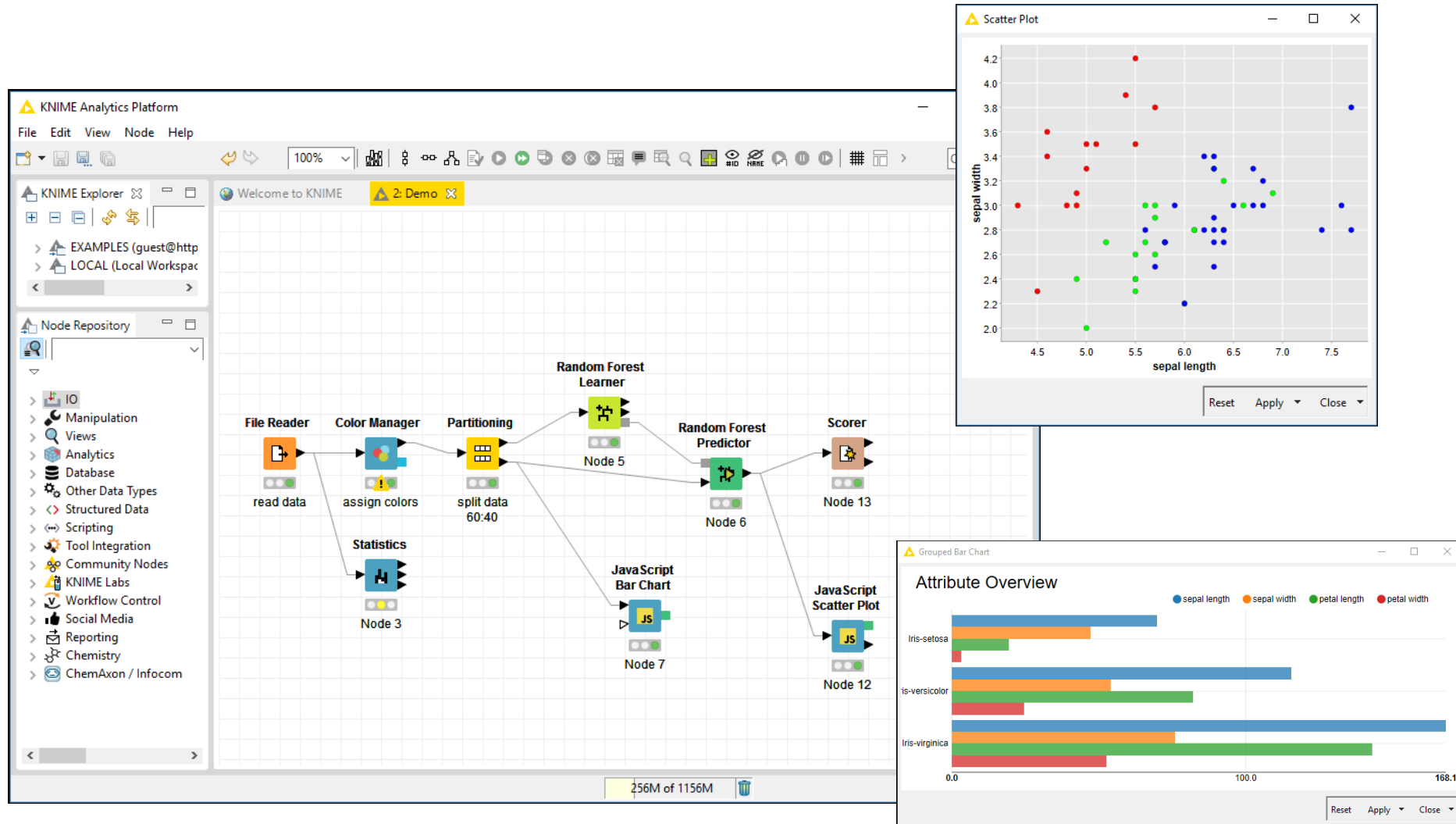
- **SOLUTIONS** for example workflows
- RESOURCES/**LEARNING HUB** www.knime.org/learning-hub
- RESOURCES/**NODE GUIDE** <https://www.knime.org/nodeguide>

KNIME Tech pages (tech.knime.org)

- **FORUM** for questions and answers
- **DOCUMENTATION** for docs, FAQ, changelogs, ...
- **COMMUNITY CONTRIBUTIONS** for dev instructions and third party nodes

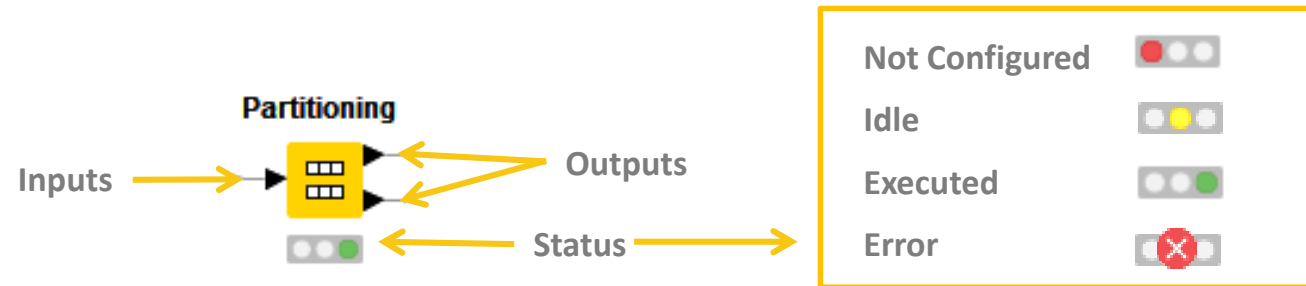
KNIME TV on YouTube <https://www.youtube.com/user/KNIMETV>

The KNIME® Analytics Platform

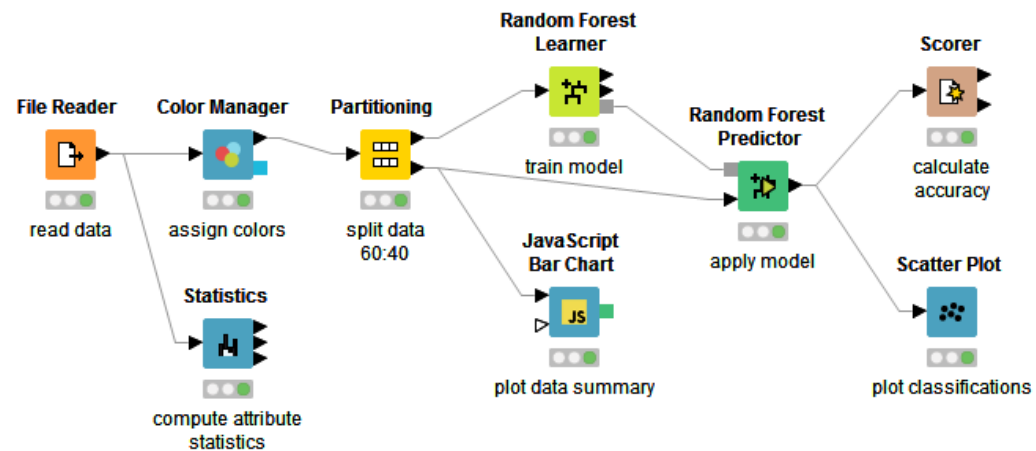


Visual KNIME Workflows

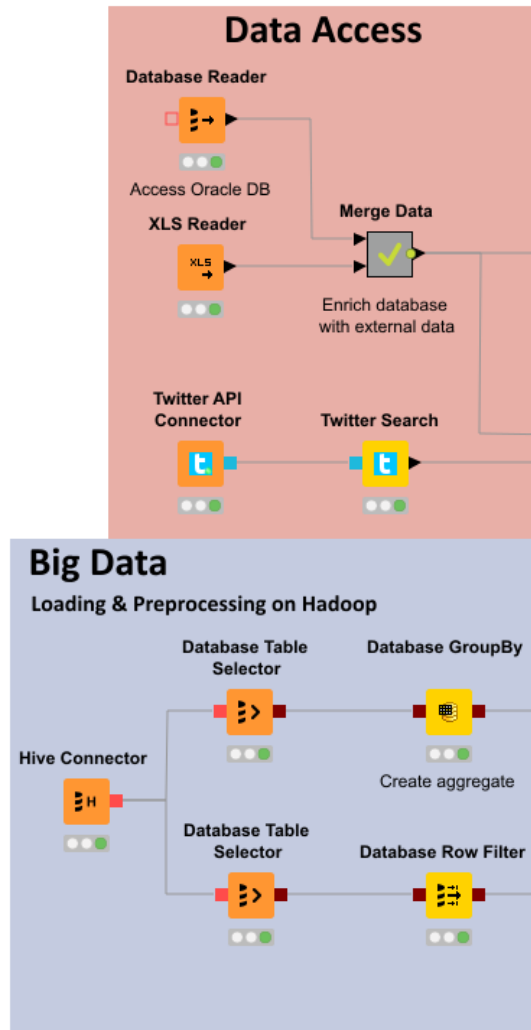
NODES perform tasks on data



Nodes are combined to create
WORKFLOWS

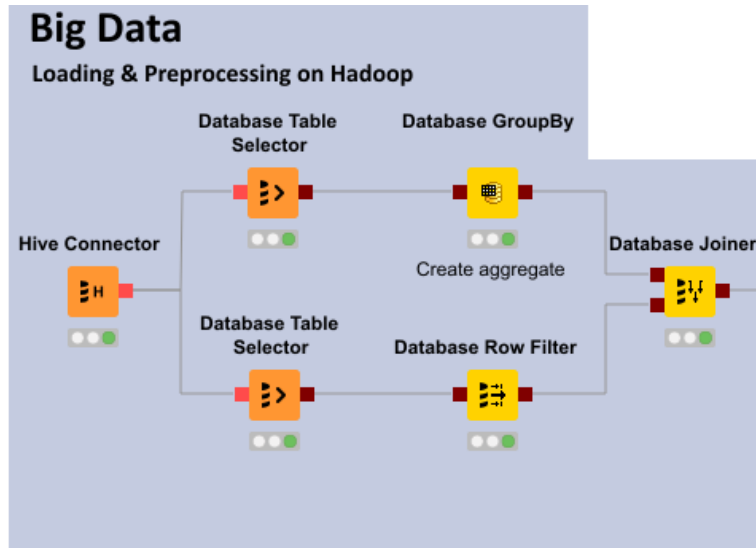


Data Access



- Databases
 - MySQL, PostgreSQL
 - any JDBC (Oracle, DB2, MS SQL Server)
- Files
 - Csv, txt
 - Excel, Word, PDF
 - SAS, SPSS
 - XML
 - PMML
 - Images, texts, networks, chem
- Web, Cloud
 - REST, Web services
 - Twitter, Google

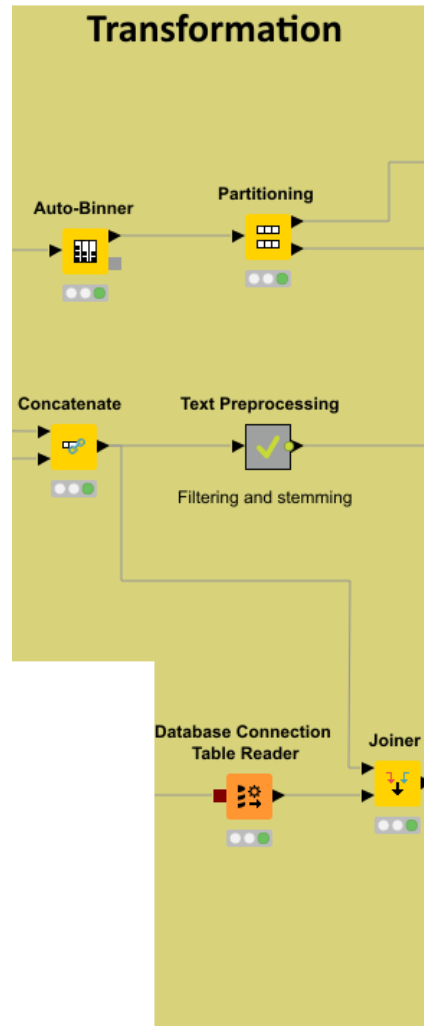
Big Data



- Spark
- HDFS support
- Hive
- Impala
- HP Vertica
- In-database processing

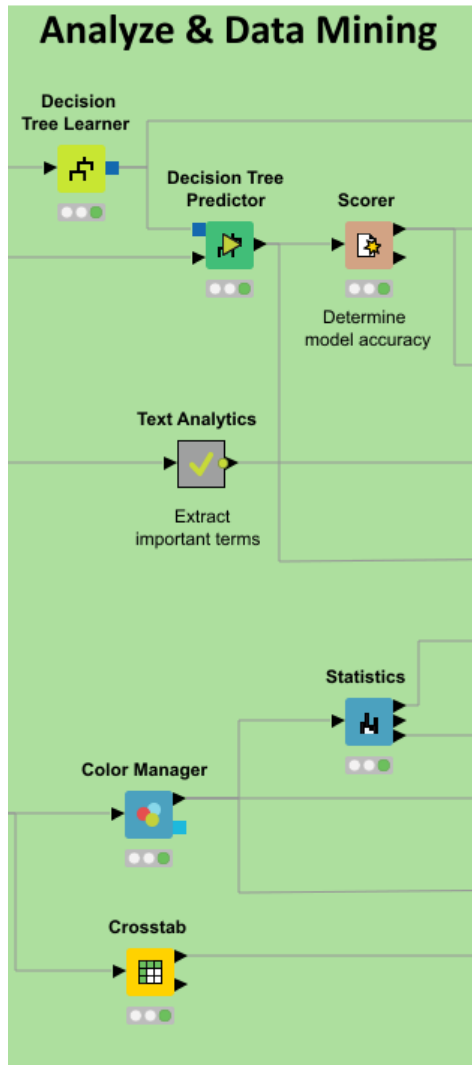


Transformation



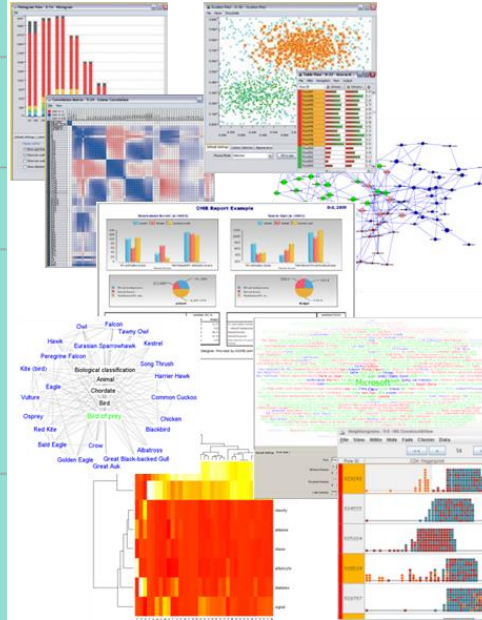
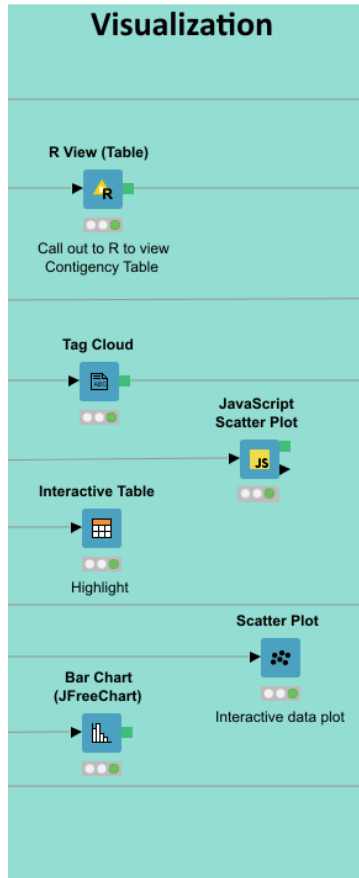
- Preprocessing
 - Row, column, matrix based
- Data blending
 - Join, concatenate, append
- Aggregation
 - Grouping, pivoting, binning
- Feature Creation and Selection

Analyze & Data Mining



- Regression
 - Linear, logistic
- Classification
 - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
 - k-means, DBSCAN, hierarchical
- Validation
 - Cross-validation, scoring, ROC
- Misc
 - PCA, MDS, item set mining
- External
 - R, Weka

Visualization



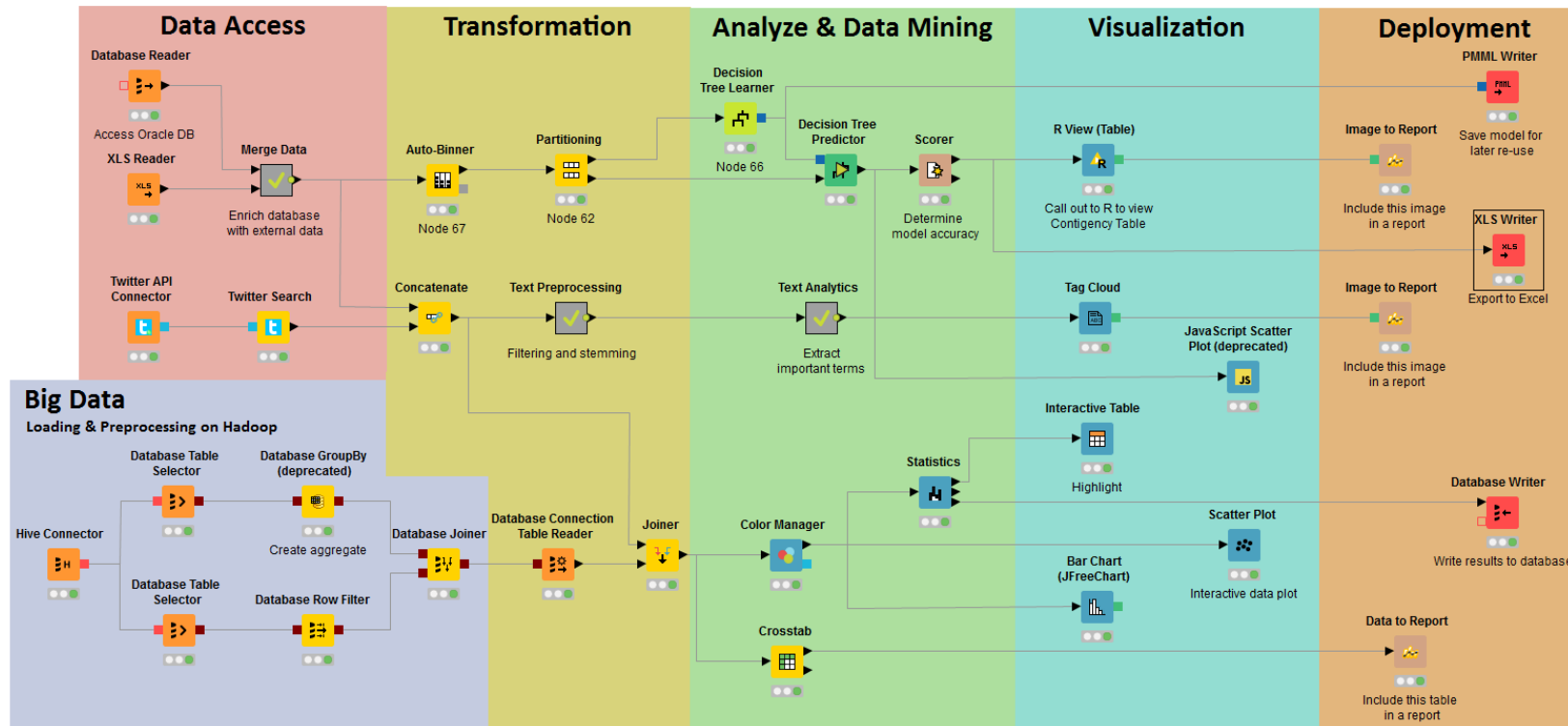
- Interactive
 - Scatter plot, histogram, pie charts, box plot
 - Highlighting (brushing)
- JFreeChart
- JavaScript
- Misc
 - Tag cloud, open street map, networks, molecules
- External
 - R

Deployment



- Database
- Files
 - Excel, csv, txt
 - XML
 - PMML
 - to: local, KNIME Server, SSH-, FTP-Server
- BIRT Reporting

Over 1500 native and embedded nodes included:



Data Access

MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers
Industry Specific
Community / 3rd

Transformation

Row,
Column
Matrix
Text, Image
Time Series
Java
Python
Community / 3rd

Analysis & Mining

Statistics
Data Mining
Machine Learning
Web Analytics
Text Mining
Network Analysis
Social Media Analysis
R, Weka, Python
Community / 3rd

Visualization

R
JFreeChart
JavaScript
Community / 3rd

Deployment

via BIRT
PMML
XML, JSON
Databases
Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd

Overview

- Installing KNIME Analytics Platform
- The KNIME Workspace
- The KNIME File Extensions
- The KNIME Workbench
 - Workflow editor
 - Explorer
 - Node repository
 - Node description
 - Preferences
- Installing new features

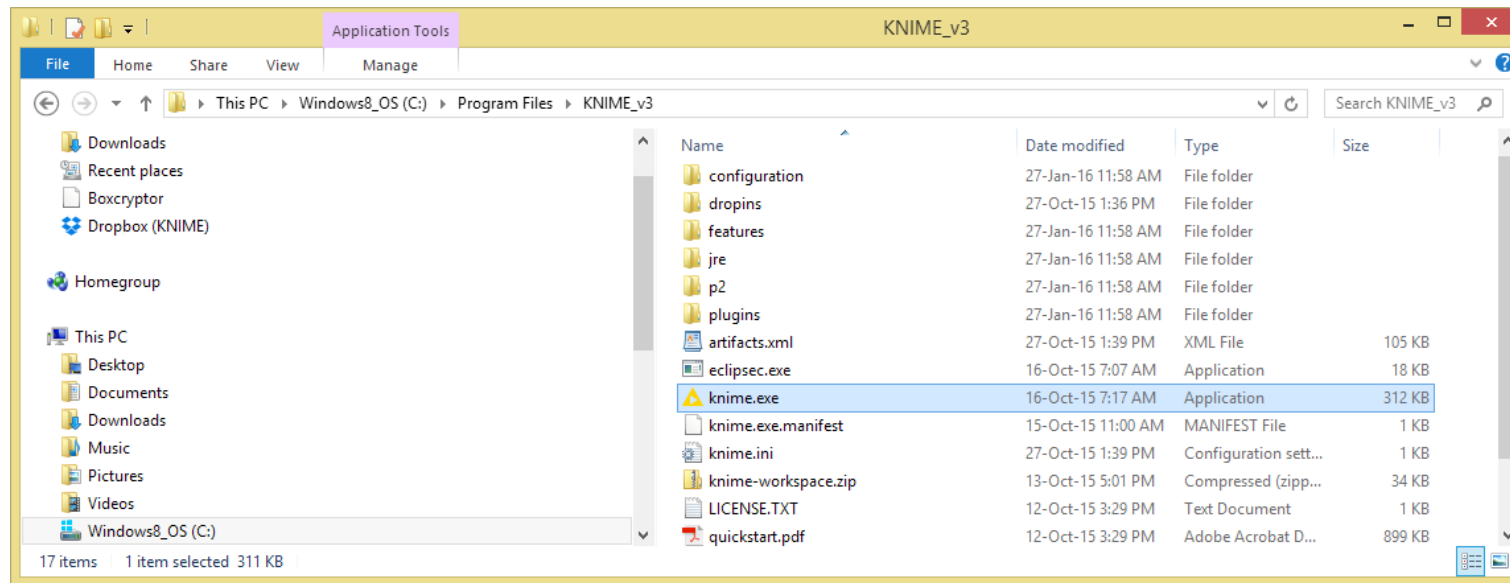
Install KNIME Analytics Platform

- Select the KNIME version for your computer:
 - Mac, Win, or Linux and 32 / 64bit
- Note different downloads (minimal or **full**)
- Download archive and extract the file, or download installer package and run it

Windows	
KNIME Analytics Platform for Windows (installer)	32 bit (301,82 MB) 64 bit (305 MB)
<i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	
KNIME Analytics Platform + all free extensions for Windows (installer)	32 bit (1,74 GB) 64 bit (1,92 GB)
<i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	
KNIME Analytics Platform for Windows (self-extracting archive)	32 bit (304,18 MB) 64 bit (306,45 MB)
<i>The self-extracting archive only creates a folder holding the KNIME installation</i>	
KNIME Analytics Platform for Windows (zip archive)	32 bit (346,61 MB) 64 bit (349,93 MB)
Linux	
KNIME Analytics Platform for Linux	32 bit (363,02 MB) 64 bit (360,1 MB)
KNIME Analytics Platform + all free extensions for Linux	32 bit (1,66 GB) 64 bit (2,07 GB)
Mac OSX	
KNIME Analytics Platform for Mac OSX (10.7 and above)	64 bit (329,41 MB)
KNIME Analytics Platform + all free extensions for Mac OSX (10.7 and above)	64 bit (1,99 GB)

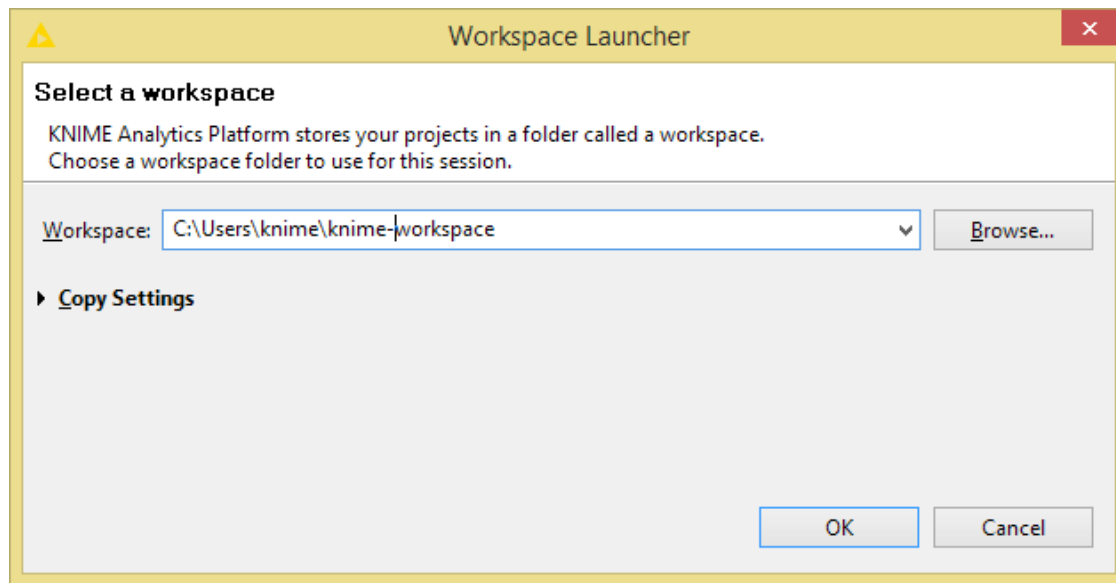
Start KNIME Analytics Platform

- Go to the installation directory and launch **KNIME**, or use the shortcut created on your Desktop.



The KNIME Workspace

- The workspace is the **folder/directory** in which workflows (and potentially data files) are stored for the current KNIME session.
- Workspaces are portable (just like KNIME)



Welcome Page



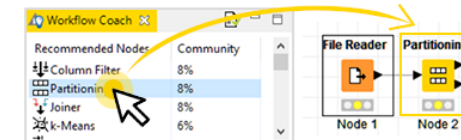
Welcome to KNIME Analytics Platform!

New to KNIME? Looking for resources to get started?

- Register for emails with introductory tips [here](#).
- Explore our [Quickstart Guide](#).
- Check out [7 things to do after installing KNIME Analytics Platform](#).
- Find more hints and how-tos in the [Learning Hub](#).
- And register for our release and event emails right [here](#).

This page will be displayed upon startup but you can customize the content using the checkboxes at the bottom.

NEW since 3.2: Workflow Coach recommends matching nodes.



Updates for the following components are available:

- DYMATRIX Uplift Modeling Extensions
- Palladian for KNIME

Click [here](#) in order to install updates.

Where to go from here

- [Create new workflow](#)
- [Learning Hub](#)
- [Browse example workflows](#)
- [Get additional nodes](#)
- [Go to my workflows](#)
- [Mount KNIME Cloud Server](#)

Most recently used workflows

- ModelSelection_WebPortal_Part1
- ModelSelection_WebPortal_Part1
- ModelSelection_BasicWorkflow
- DataCleaning_WebPortal_v2.0
- KNIME_project2
- Sexy ETL_v2.0

Tips & Tricks

Specialist Nodes

Did you know there are a whole variety of specialist nodes available from KNIME Labs and the Community around Scripting, Image Processing, Text Processing, Internet Mining, Network Mining, Cell Biology and Genetics, and Chemistry. To access them, go to Help Menu and choose Install New< Software.

- ☒ Show intro text at next start
- ☒ Show update notifications at next start
- ☒ Show links and most recently used workflows at next start

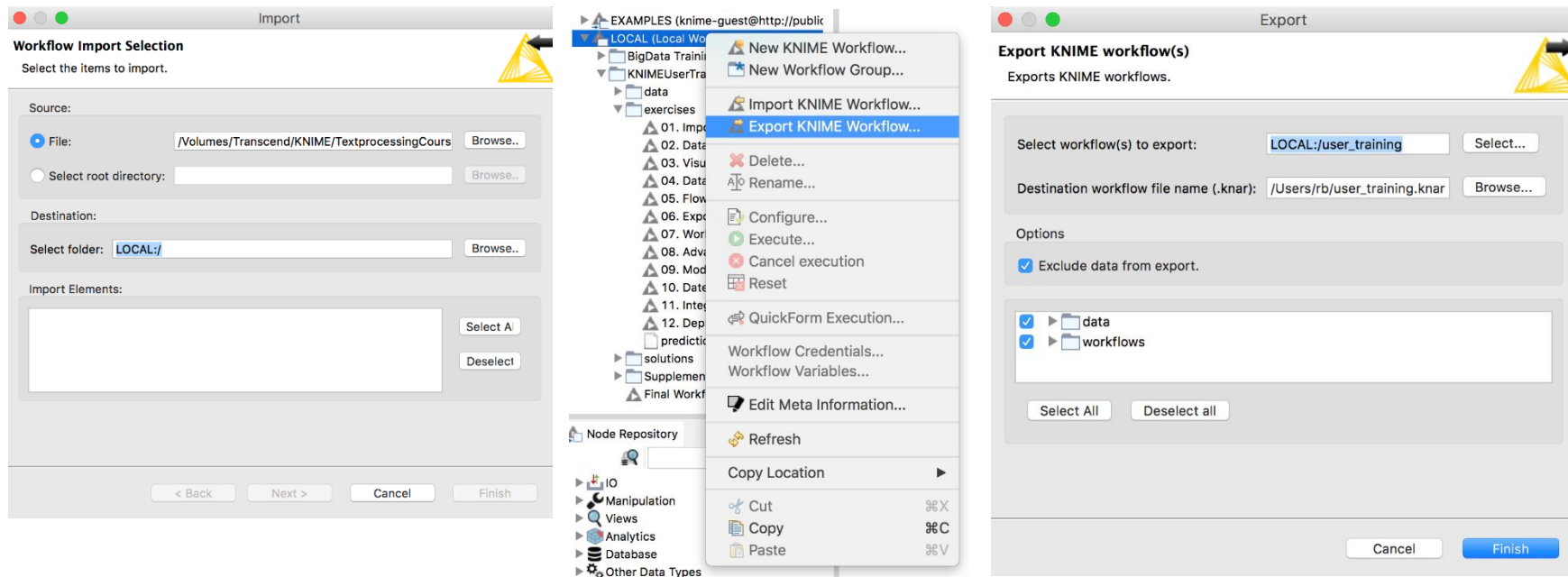
The KNIME Workbench

The screenshot displays the KNIME Analytics Platform interface with several key components highlighted by yellow callout boxes:

- Servers and Workflows:** Points to the **KNIME Explorer** on the left, which shows a project tree with folders like **EXAMPLES**, **LOCAL (Local Workspace)**, **BigData Training**, **KNIMEUserTraining**, **data**, **exercises**, and **solutions**. A specific workflow, **04. Data Mining - solution**, is selected.
- Workflow Editor:** Points to the central workspace where a workflow is being built. The workflow starts with **Fully Joined Data**, followed by a **Partitioning** node, then a **Decision Tree Learner** node, a **Decision Tree Predictor** node, and finally a **Scorer** node. A **JavaScript ROC Curve** node and a **Node 278 ROC Curve** node are also visible.
- Node Recommendations:** Points to the **Recommended Nodes** panel on the left, which lists nodes like **Decision Tree Predictor**, **Decision Tree Learner**, **Naive Bayes Predictor**, **PIRI Naive Bayes Learner**, and **Regression Predictor**.
- Node Repository:** Points to the **Node Repository** panel on the left, which shows a hierarchical view of node categories such as **IO**, **Manipulation**, **Views**, **Analytics**, **Database**, **Other Data Types**, **Structured Data**, **Scripting**, **Tool Integration**, **Community Nodes**, and **KNIME Labs**.
- Node Description:** Points to the **Node Description** panel on the right, which provides detailed information about the selected **Partitioning** node. The description includes a brief overview, **Dialog Options** (Absolute, Relative, Top from top, Linear sampling), and a **Console** output area.
- Outline:** Points to the **Outline** panel at the bottom left, which shows a hierarchical view of the workflow nodes.
- Console:** Points to the **Console** panel at the bottom right, which displays the KNIME Console output, including a welcome message and license information.

Creating New Workflows, Importing and Exporting

- Right-click Workspace in KNIME Explorer to create new workflow or workflow group or to import workflow
- Right-click on workflow or workflow group to export



KNIME File Extensions

- Dedicated file extensions for Workflows and Workflow groups associated with KNIME Analytics Platform

- ***.knwf** for KNIME Workflow Files



- ***.knar** for KNIME Archive Files



More on Nodes...

A node can have 3 states:

File Reader



Idle:

The node is not yet configured and cannot be executed with its current settings.

File Reader



Configured:

The node has been set up correctly, and may be executed at any time

File Reader

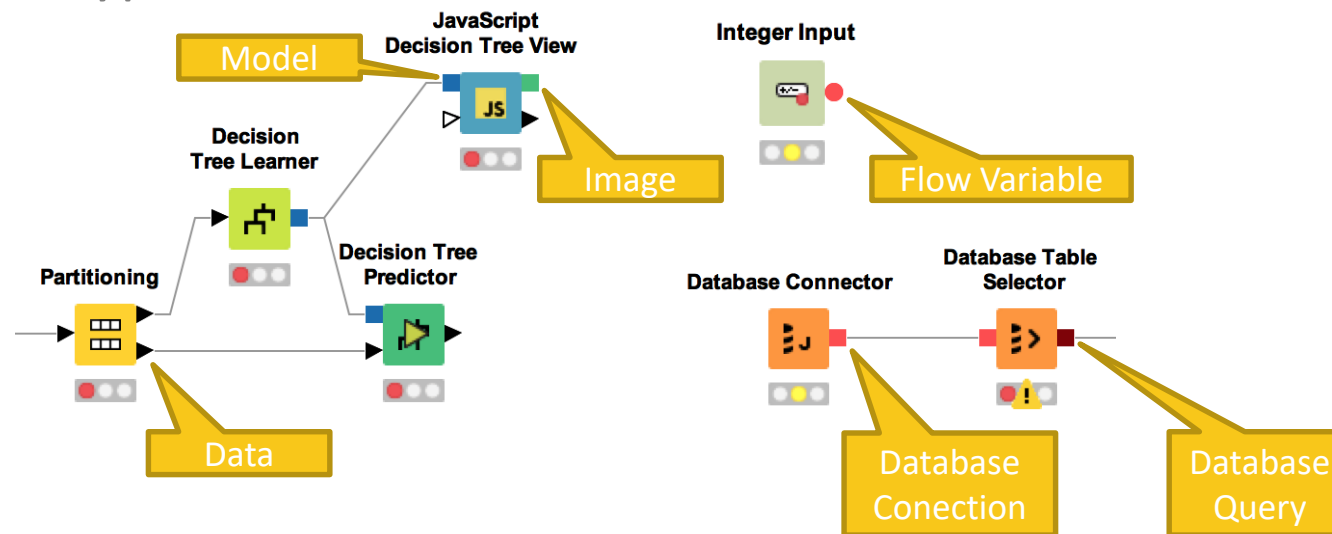


Executed:

The node has been successfully executed. Results may be viewed and used in downstream nodes.

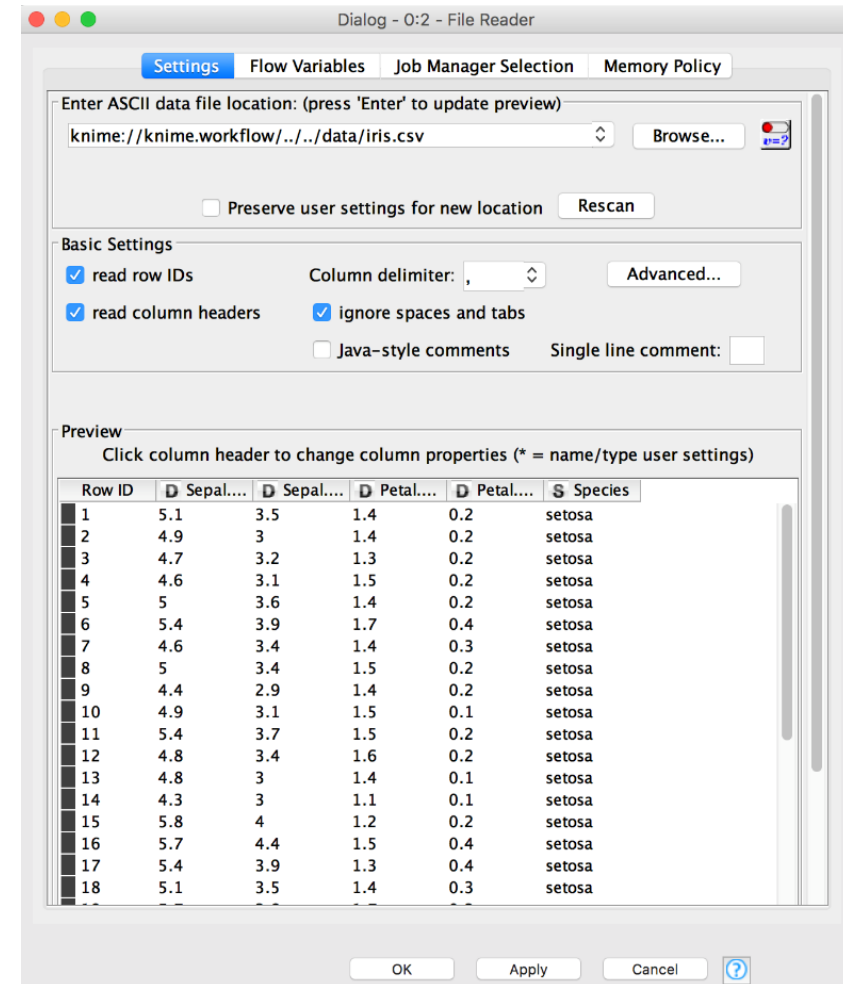
Inserting and Connecting Nodes

- Insert nodes into workspace by dragging them from Node Repository or by double-clicking in Node Repository
- Connect nodes by left-clicking output port of Node A and dragging the cursor to (matching) input port of Node B
- Common port types:



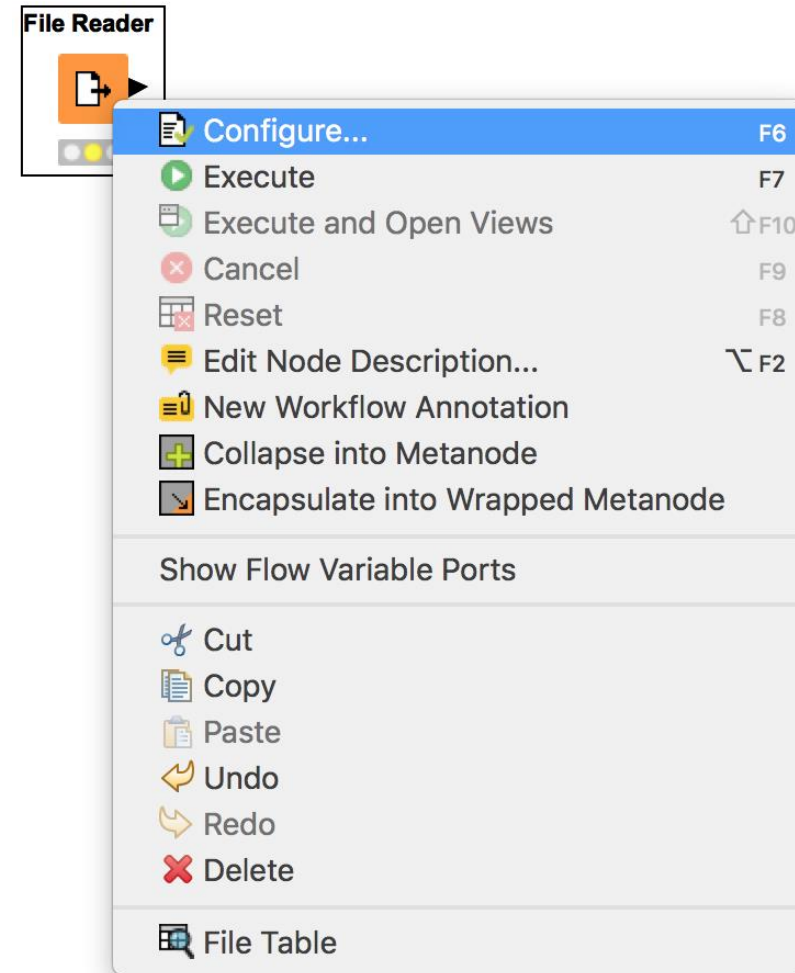
Node Configuration

- Most nodes require configuration
- To access a node configuration window:
 - Double-click the node
 - Right-click > Configure



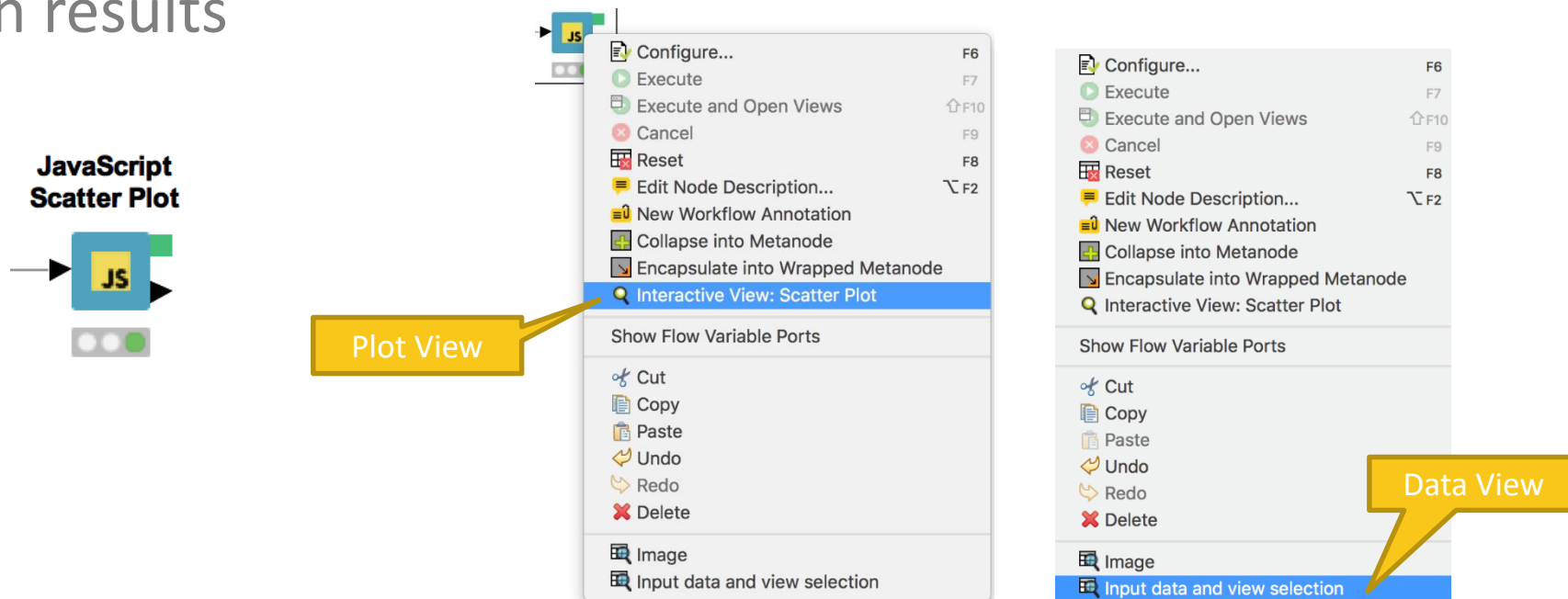
Node Execution

- Right-click node
- Select Execute in context menu
- If execution is successful, status shows green light
- If execution encounters errors, status shows red light



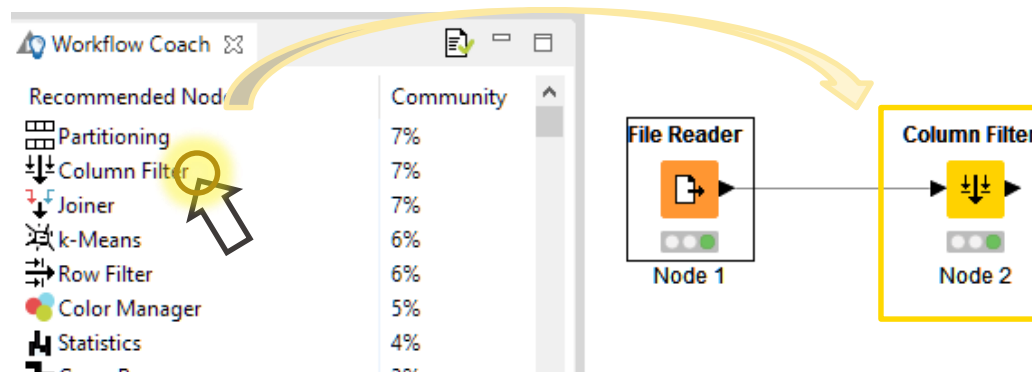
Node Views

- Right-click node
- Select Views in context menu
- Select output port to inspect execution results



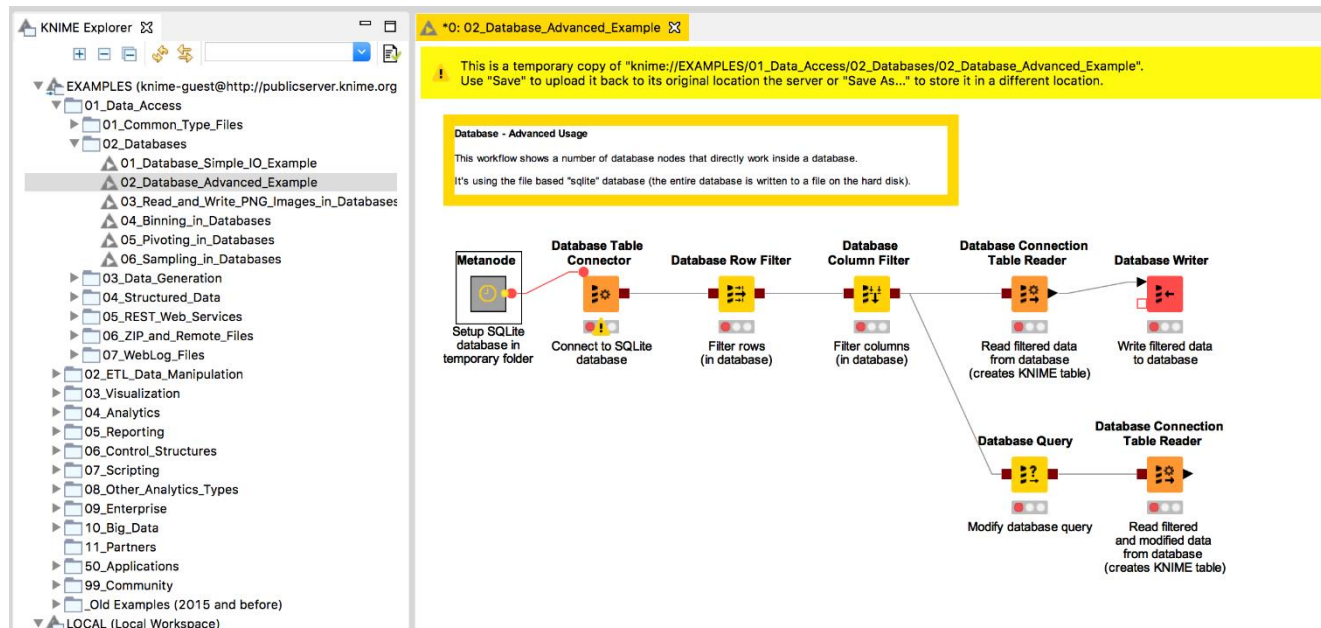
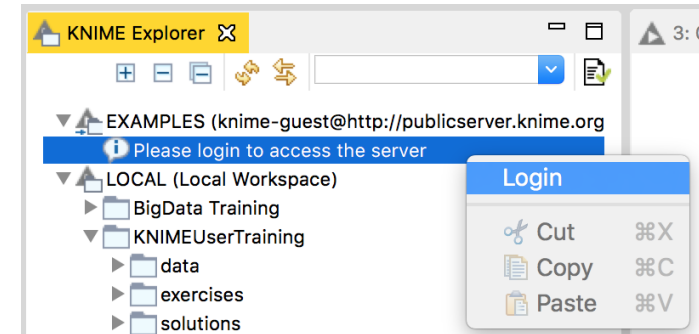
Workflow Coach

- Recommendation engine
 - It gives hints about which node use next in the workflow
 - Based on KNIME communities' usage statistics
 - Usage statistics available also with Personal Productivity Extension and KNIME Server products (these products require a purchased license)



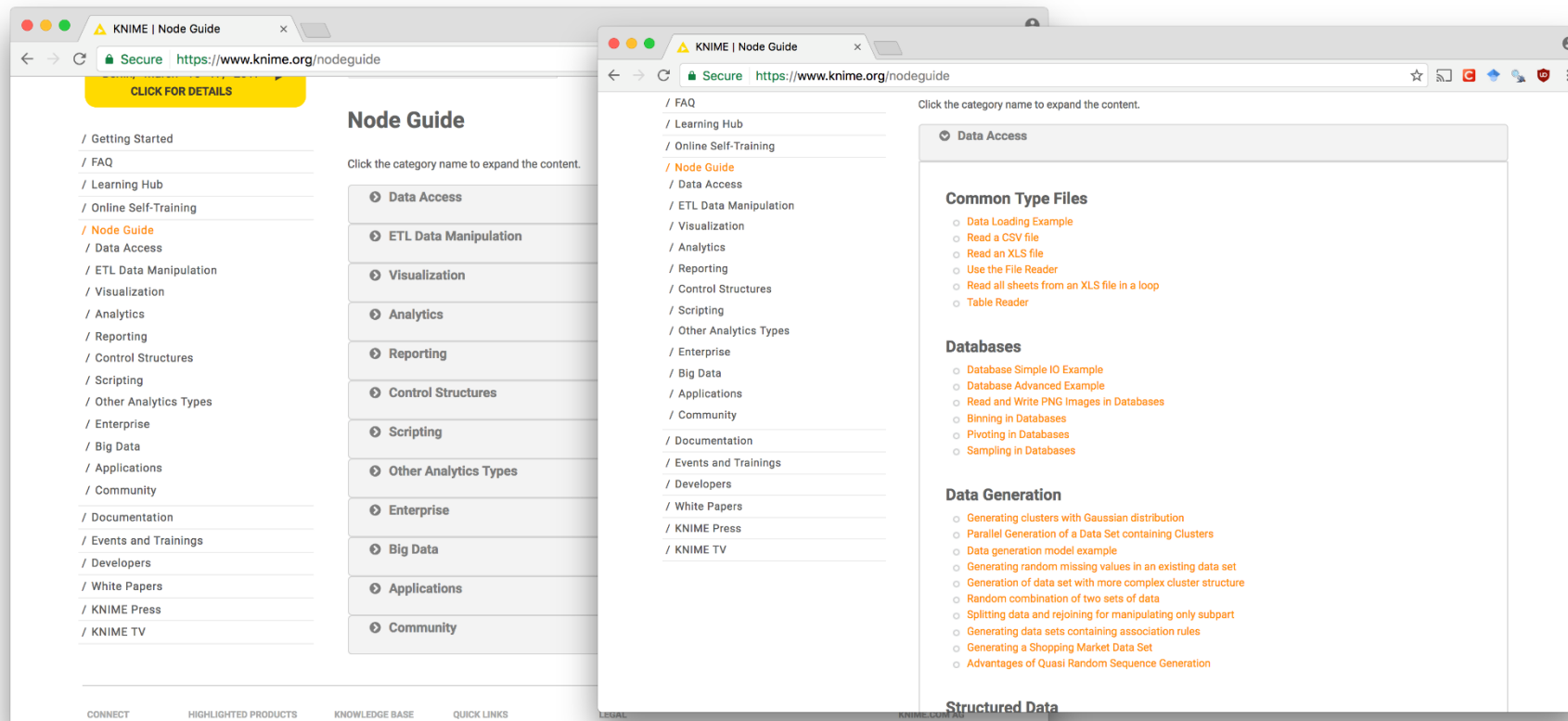
Getting Started: KNIME Example Server

- Public repository with large selection of example workflows for many, many applications
- Connect via KNIME Explorer



Online Node Guide

- Workflows from Example Server also available online
 - <https://www.knime.org/nodeguide>



Hot Keys (for future reference)

Task	Hot key	Description
Node Configuration	F6	opens the configuration window of the selected node
Node Execution	F7	executes selected configured nodes
	Shift + F7	executes all configured nodes
	Shift + F10	executes all configured nodes and opens all views
	F9	Cancels selected running nodes
	Shift + F9	Cancels all running nodes
Move Nodes and Annotations	Ctrl + Shift + Arrow	moves the selected node in the arrow direction
	Ctrl + Shift + PgUp/PgDown	moves the selected annotation in the front or in the back of all overlapping annotations
Workflow Operations	F8	resets selected nodes
	Ctrl + S	saves the workflow
	Ctrl + Shift + S	saves all open workflows
	Ctrl + Shift + W	closes all open workflows
Meta-node	Shift + F12	opens meta-node wizard

Introduction to the Big Data Course

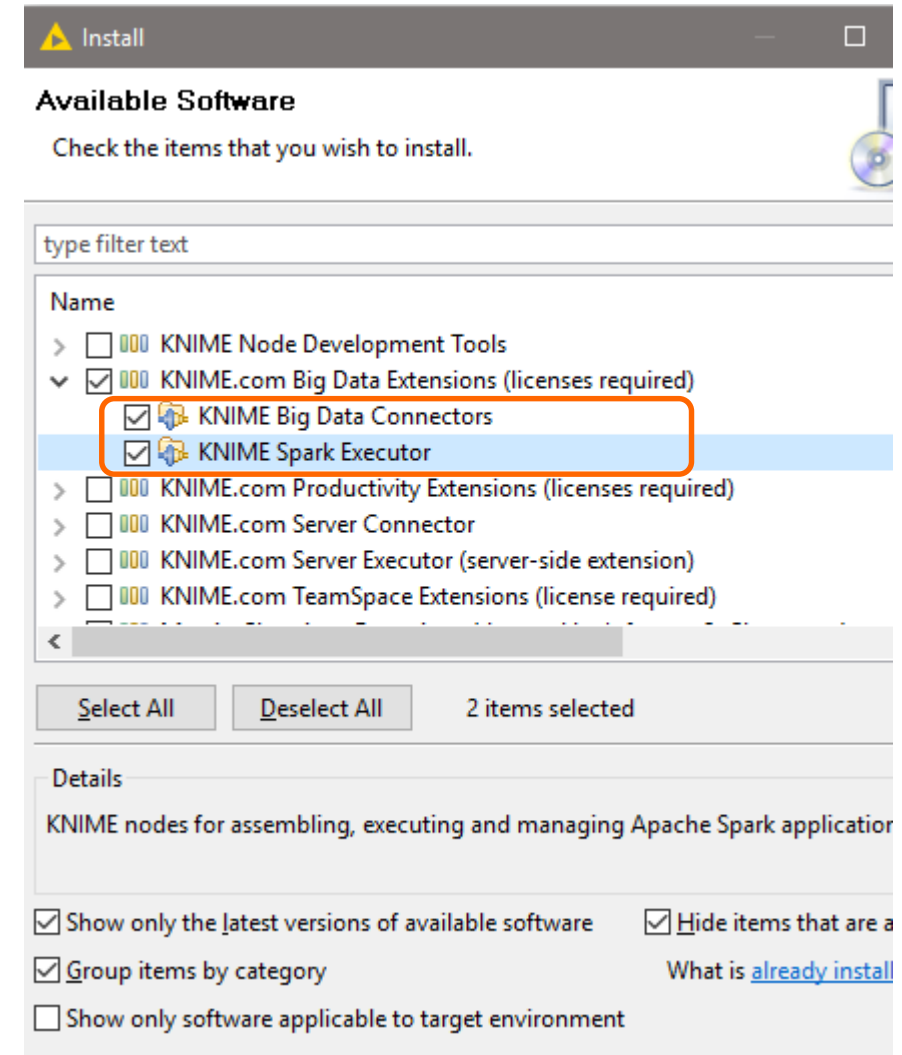
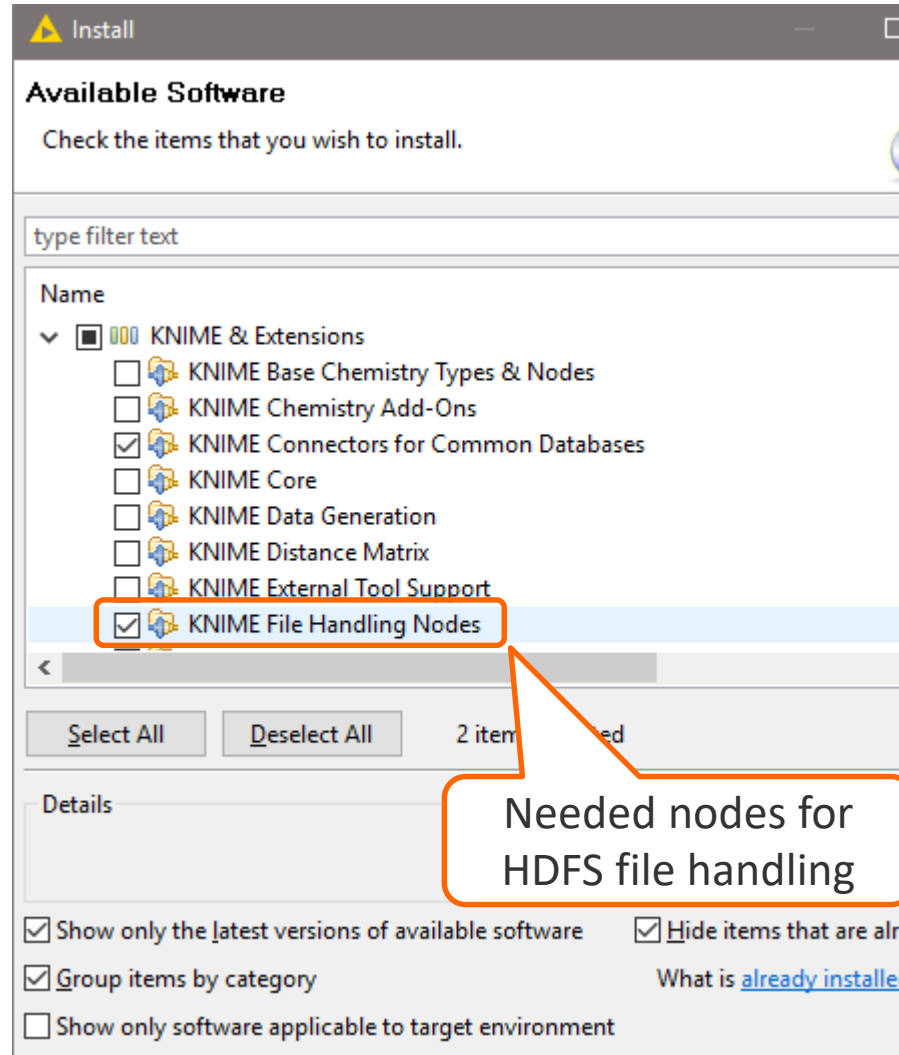
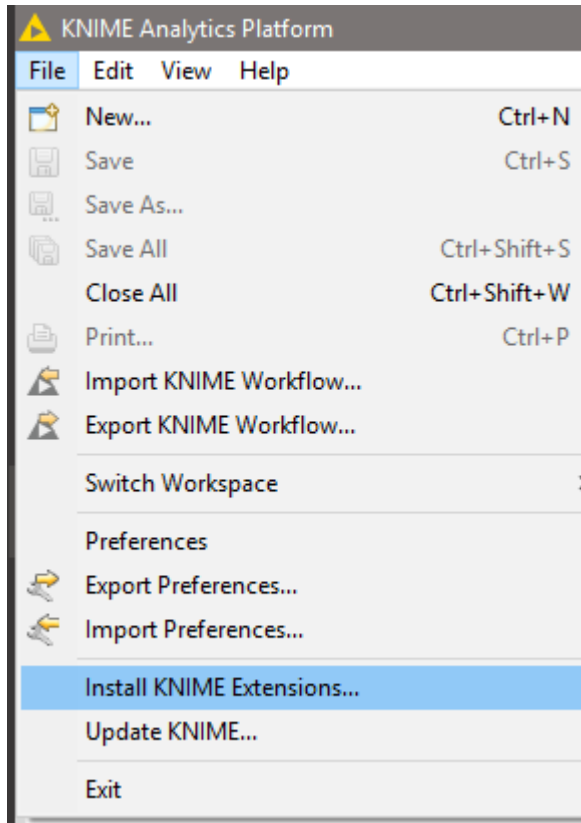
Goal of this Course

Become familiar with the **KNIME Big Data Extensions** to operate on **Hadoop** and **Spark** based platforms.

What you need:

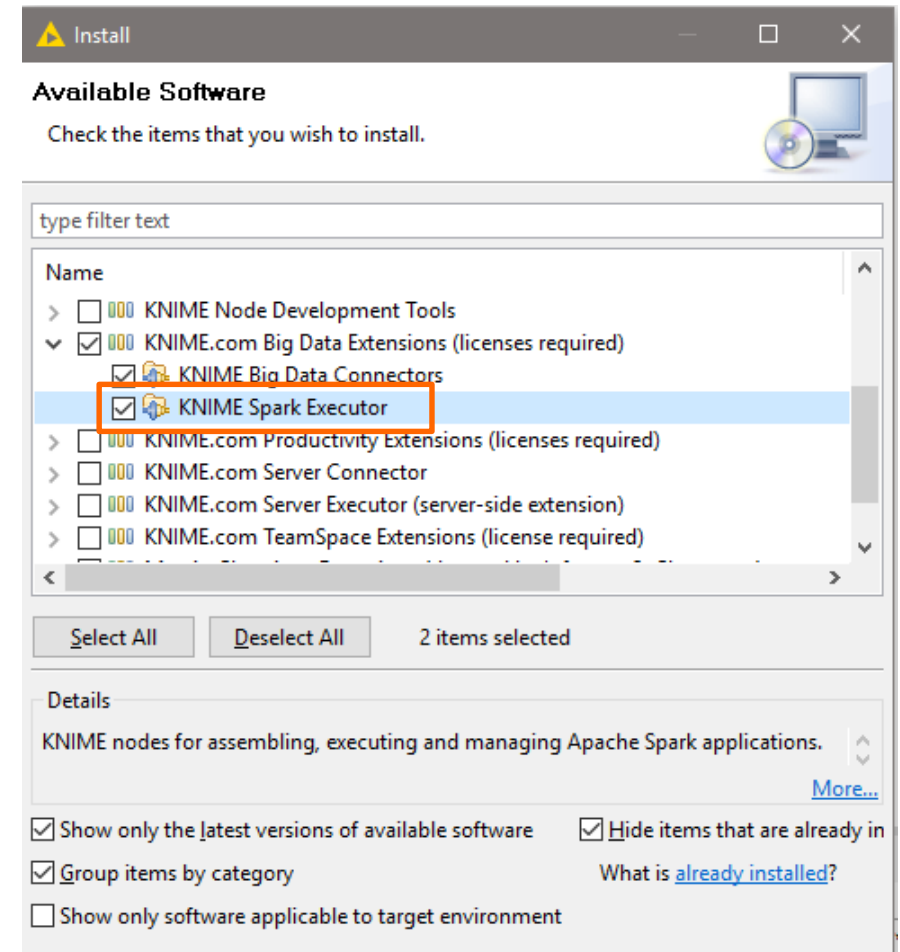
- Install KNIME Big Data Extensions
 - Big Data Connectors
 - Spark Executor
- Big Data License (in USB stick valid 1 week - complimentary)

Installation of File Handling and Big Data Extensions



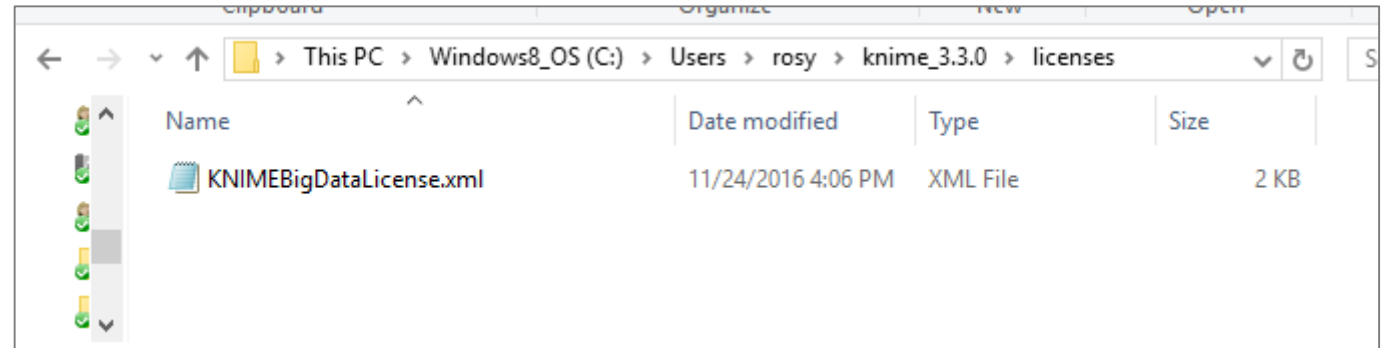
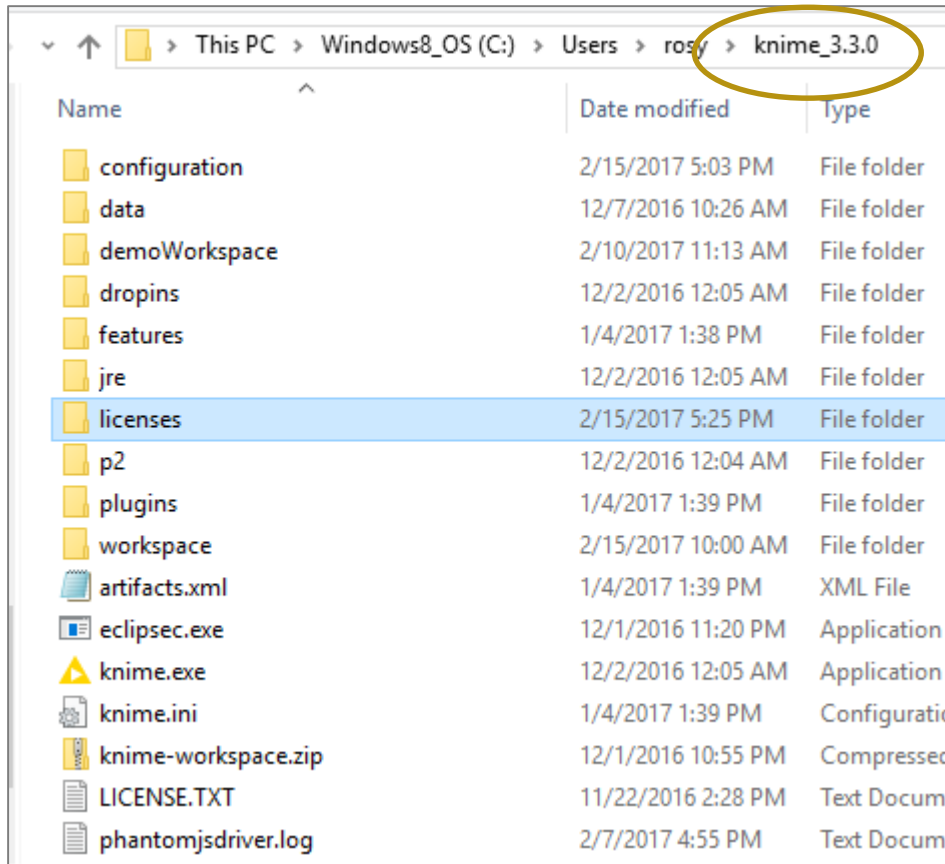
Install Spark Extension

- Supported Spark Versions
 - 1.2, 1.3, 1.5, 1.6
- One KNIME Spark Executor for all Spark versions

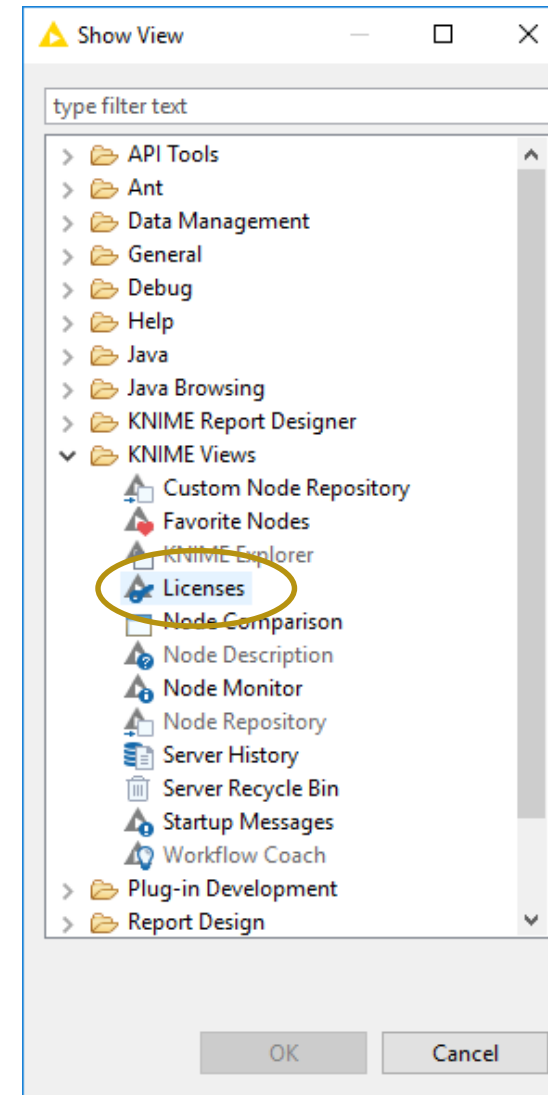
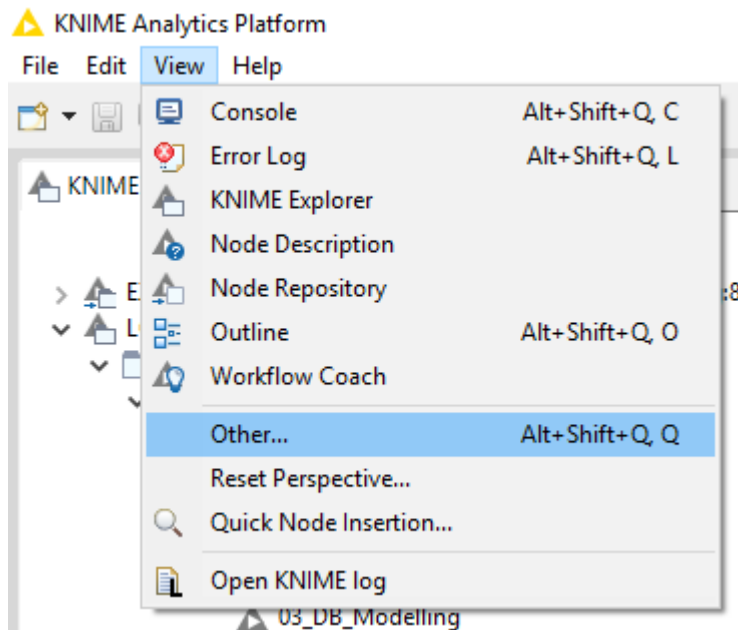


Test License

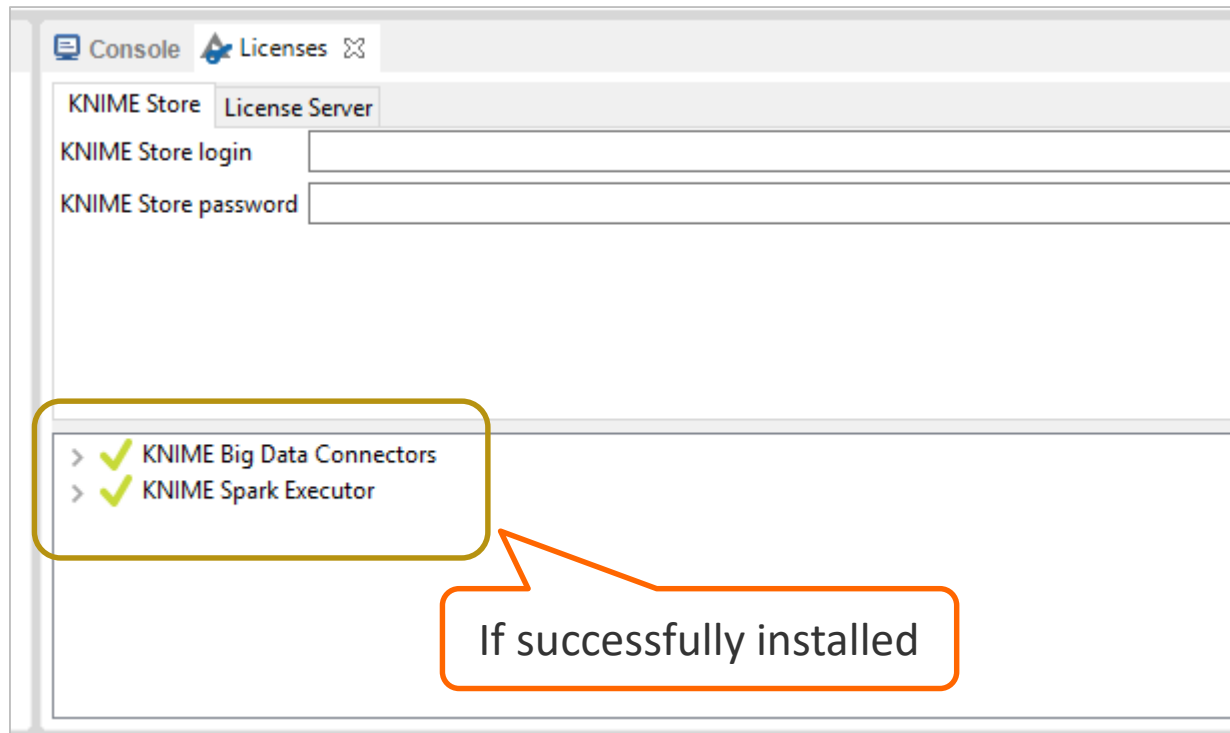
- Copy license xml file into licenses folder in KNIME Installation Folder



Monitor licenses through Licenses View



License View



- License file
KNIMEBigDataLicense.xml in
USB stick
- 2 licenses: Hadoop + Spark
- Valid 1 week (complimentary)
- 30-day free test license
<https://www.knime.org/knime-big-data-extensions-free-30-day-trial>
- Product Description
<https://www.knime.org/knime-performance-extensions>

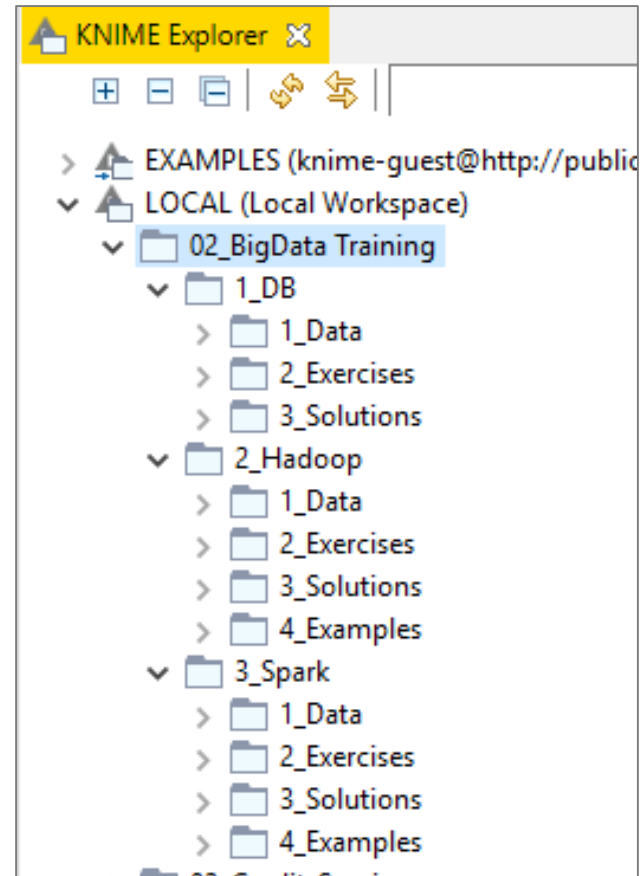
Big Data Resources (1)

- SQL Syntax and Examples www.w3schools.com
- Apache Spark MLlib <http://spark.apache.org/mllib/>
- KNIME Performance Extension (Hadoop + Spark)
<https://www.knime.org/knime-performance-extensions>
- Free 30-days test license <https://www.knime.org/knime-big-data-extensions-free-30-day-trial>

Big Data Resources (2)

- Whitepaper “KNIME opens the Doors to Big Data”
https://www.knime.org/files/big_data_in_knime_1.pdf
- Blog Posts
 - <https://www.knime.org/blog/Hadoop-Hive-meets-Excel>
 - <https://www.knime.org/blog/hive-execution-engine-comparison-with-the-knime-analytics-platform>
- Example workflows on EXAMPLES Server in 10_Big_Data

Workflows for this Course



Steps

- Problem Definition
- Problem Solution using a traditional Database, Database Nodes, and KNIME native Machine Learning Nodes
- Moving In-Database Processing from Database to Hadoop Hive Platform
- Moving In-Database Processing and Machine Learning to Spark

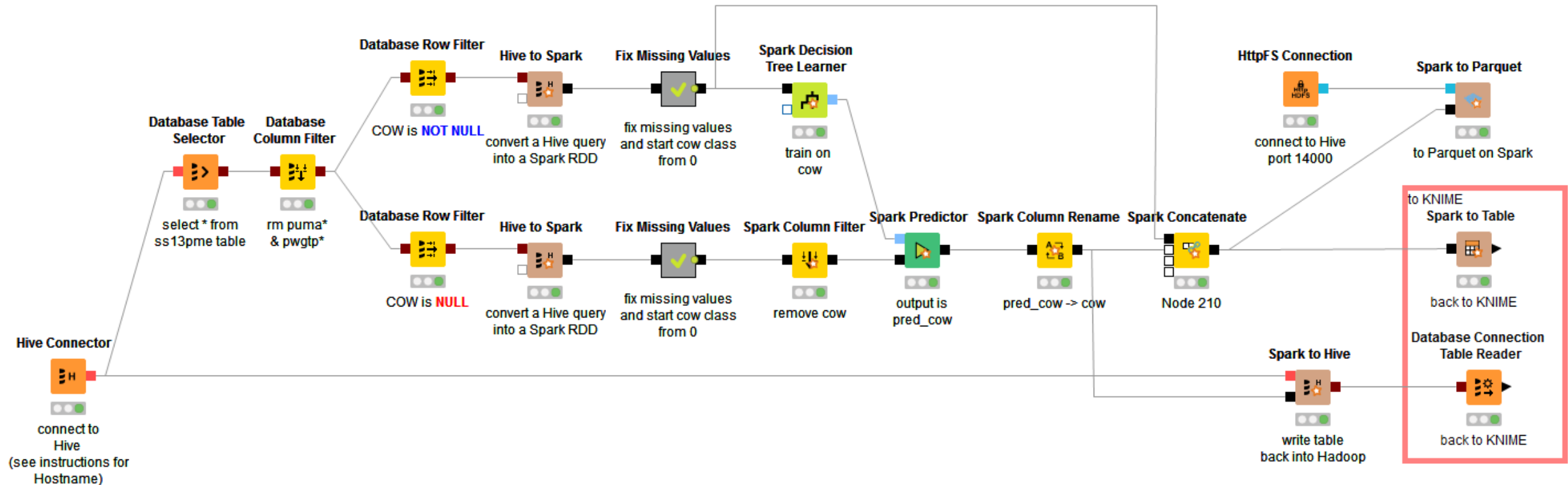
Today's Example: Missing Values Strategy

- Missing Values are a big problem in Data Science!
- Many strategies to deal with the problem (see “How to deal with missing values” KNIME Blog post of 10/21/2014
<https://www.knime.org/blog/how-to-deal-with-missing-values>)
- We adopt the strategy that predicts the missing values based on the other attributes on the same data row
- CENSUS Data Set with missing COW values from
<http://www.census.gov/programs-surveys/acs/data/pums.html>

CENSUS Data Set

- CENSUS data contains questions to a sample of US residents (1%) over 10 years
- CENSUS data set description: http://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict15.pdf
- *ss13hme* (60K rows) -> questions about housing to Maine residents
- *ss13pme* (60K rows) -> questions about themselves to Maine residents
- *ss13hus* (31M rows) -> questions about housing to all US residents in the sample
- *ss13pus* (31M rows) -> questions about themselves to all US residents in the sample

Today's Example: Missing Values Strategy



Missing Values Strategy Implementation

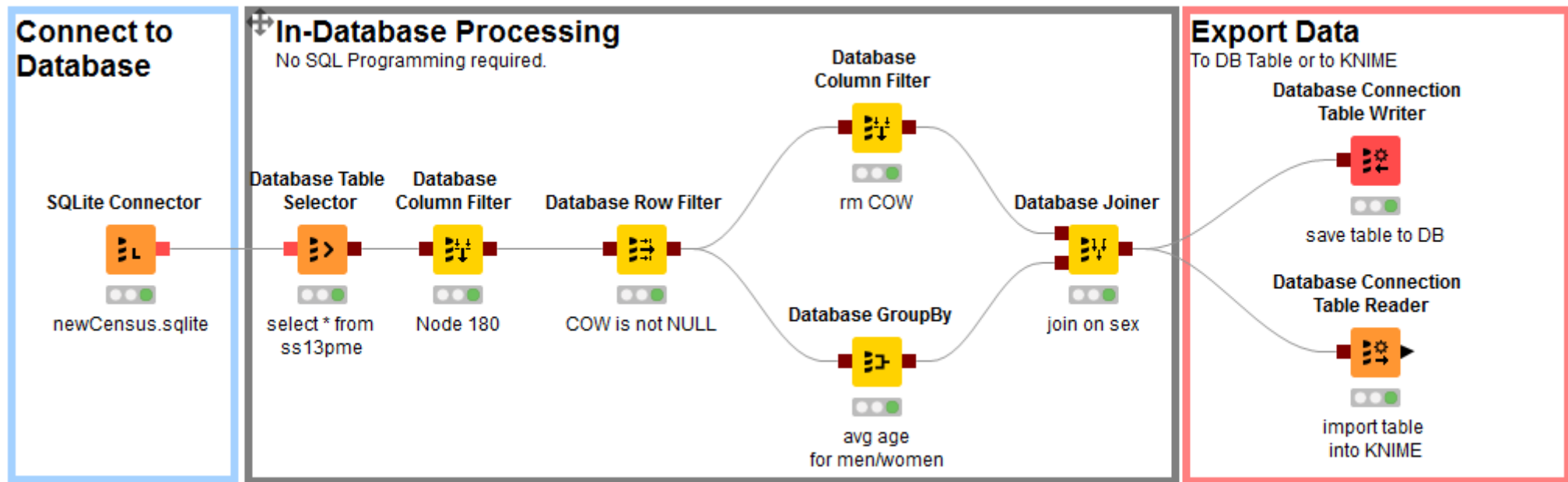
- Connect to Data (CENSUS data set)
- Aggregate and join aggregations with original data (various other ETL operations just for demo)
- Separate data rows with income from data rows with missing income
- Train a decision tree to predict income (obviously only on data rows with income)
- Apply decision tree to predict income where income is missing
- Update original data set with new predicted income values

Let's practice first on a traditional Database

Database Extension

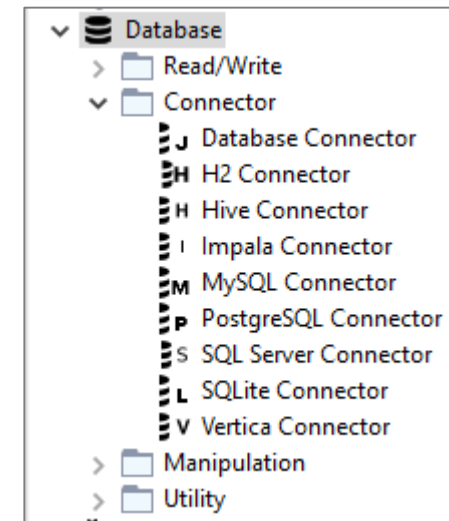
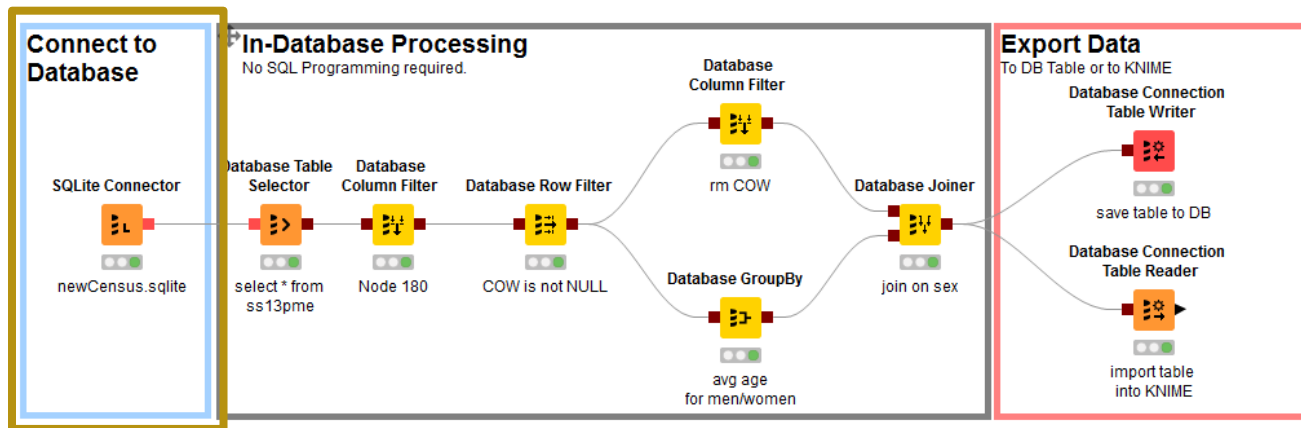
Database Extension

- Visually assemble complex SQL statements (no SQL coding needed)
- Connect to all JDBC-compliant databases
- Harness the power of your database within KNIME



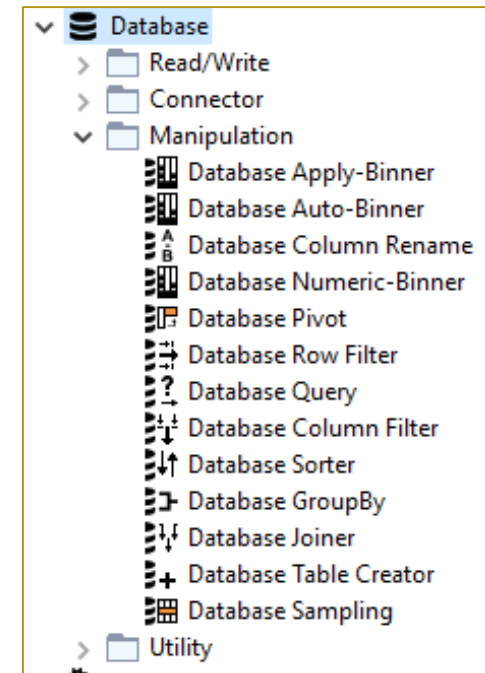
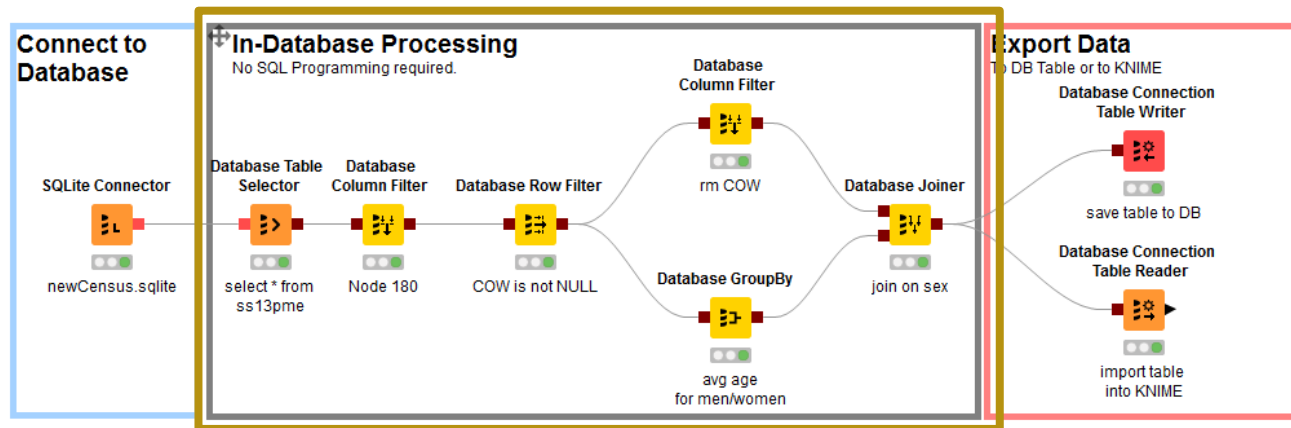
Database Connectors

- Many dedicated DB Connector nodes available
- If connector node missing, use Database Connector node with JDBC driver
- JDBC driver to upload in Preferences -> KNIME -> Databases (“Add File”)



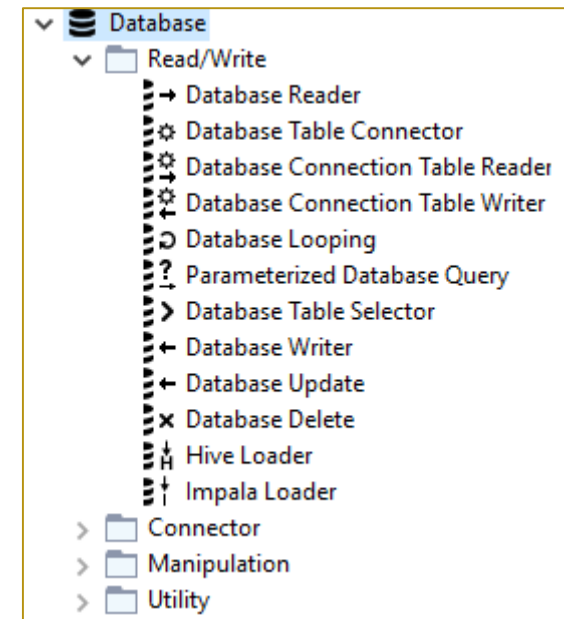
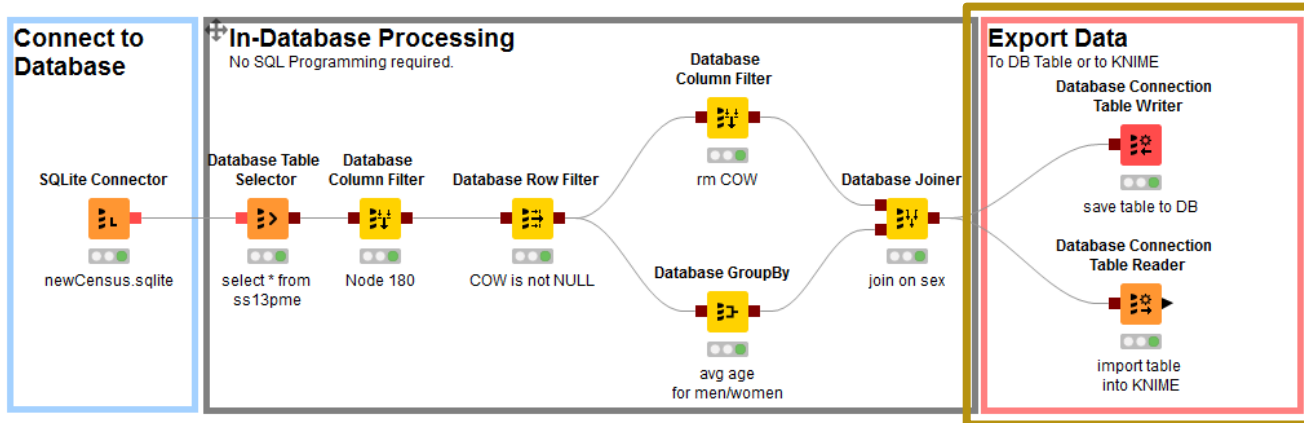
In-Database Processing

- Database Manipulation nodes generates a SQL query on top of the input SQL query (brown square port)
- Only Database Query node requires SQL code; all other Database Manipulation nodes create the SQL query for you



Export Data

- Writing data back into database
- Exporting data into KNIME
- SQL operations are **executed on the database!**



Tip

- SQL statements are logged in KNIME log file

The screenshot displays the KNIME Analytics Platform interface. The main workspace shows a workflow with several nodes: SQLite Connector, Database Table Selector, Database GroupBy, Database Joiner, Database Connection Table Writer, Database Table Selector, Database Connection Table Reader, and Decision Tree Learner. The 'Database Table Selector' node is highlighted, and its 'Node Description' panel on the right is open, showing the description and dialog options for this node.

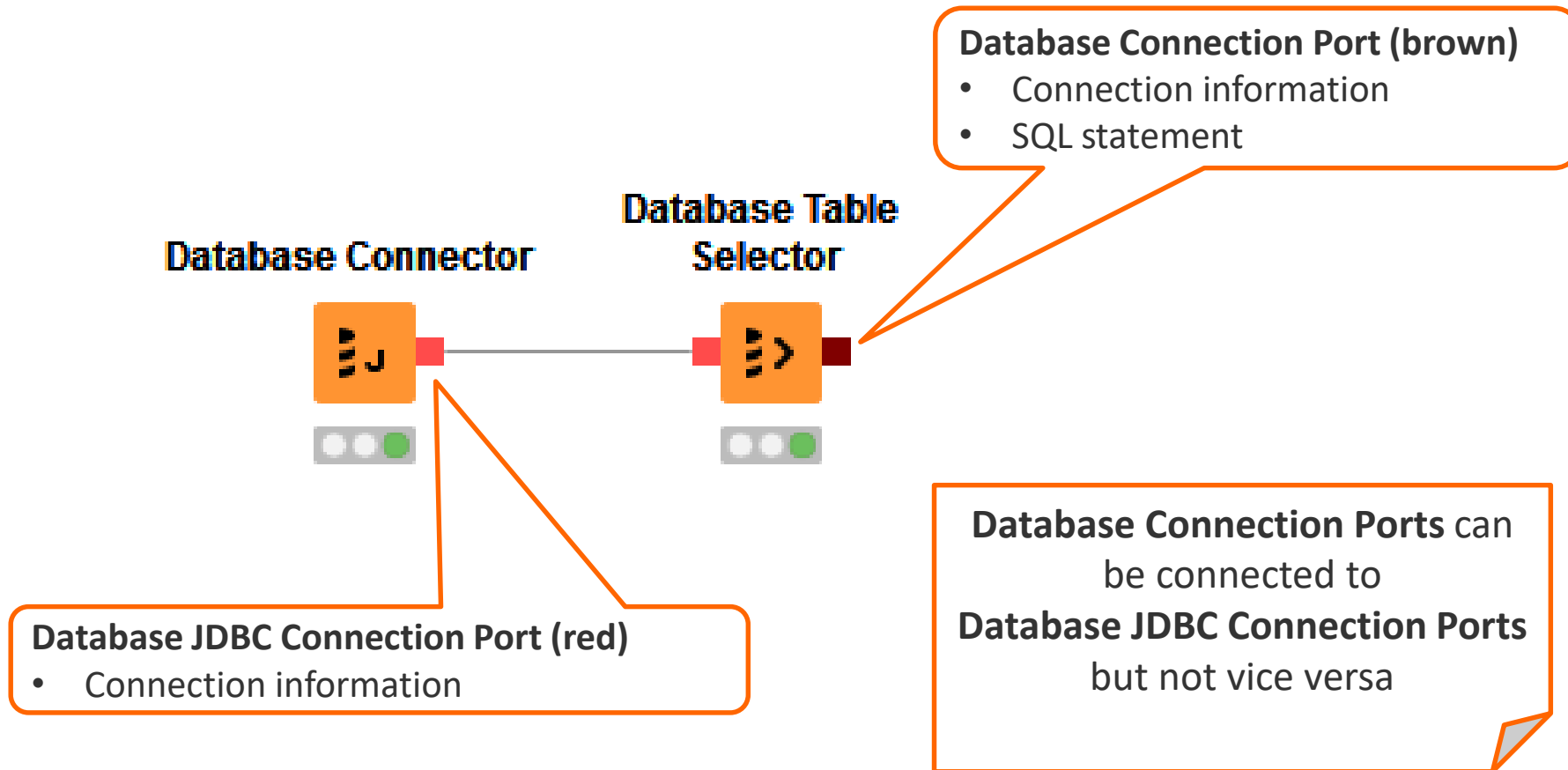
The 'Node Description' panel for 'Database Connection Table Reader' is visible, showing the description: 'Opens a database connection based on the input settings and reads the entire data table from the database into KNIME.' and the 'Dialog Options' section, which includes the option 'Use database row id'.

The 'Console' window at the bottom shows the log output. The log entry for the 'Database Table Selector' node is highlighted, showing the SQL statement executed: `SELECT * FROM model`.

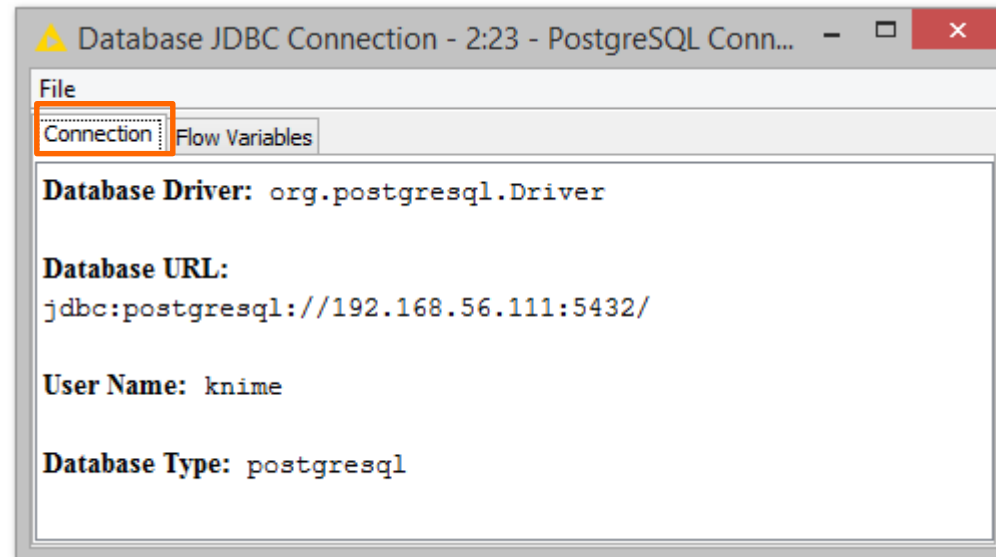
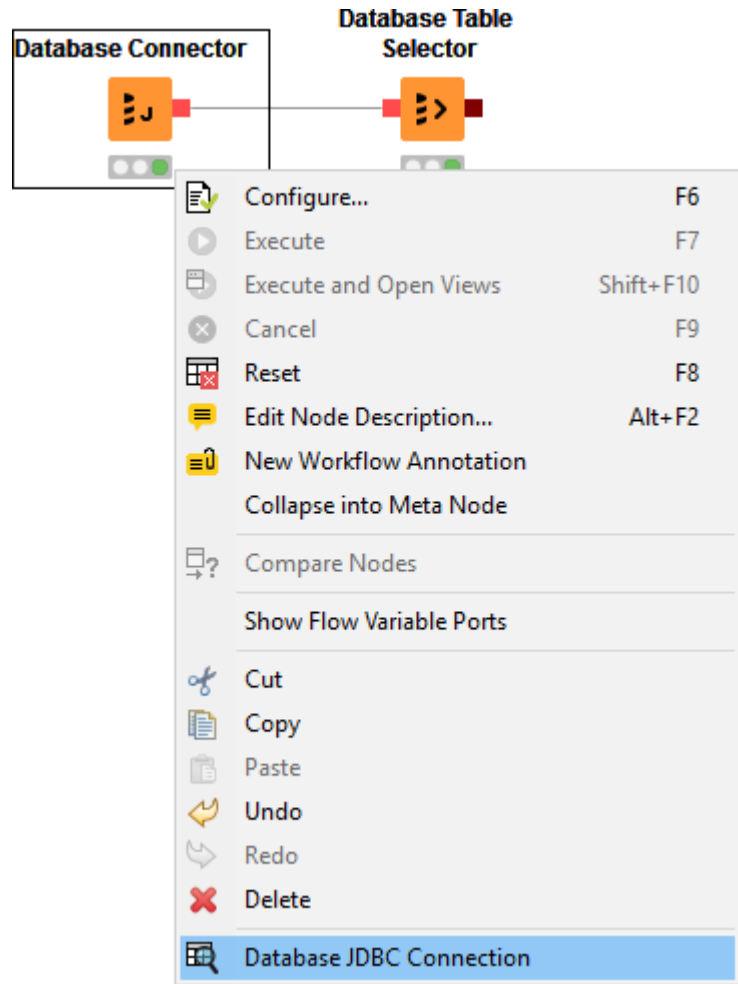
```
2016-02-01 10:10:54,371 : DEBUG : KNIME-Worker-25 : DatabaseDriverLoader : Database Table Selector : 0:188:179 : Database driver org.sqlite.JDBC retrieved from driver map
2016-02-01 10:10:54,371 : DEBUG : KNIME-Worker-25 : DBReaderImpl : Database Table Selector : 0:188:179 : Executing SQL statement as executeQuery: SELECT * FROM model
2016-02-01 10:10:54,372 : DEBUG : KNIME-Worker-25 : DBReaderImpl : Database Table Selector : 0:188:179 : Reading more data from database ResultSet
2016-02-01 10:10:54,373 : DEBUG : KNIME-Worker-25 : Database Table Selector : Database Table Selector : 0:188:179 : Configure succeeded (Database Table Selector)
```

Database Port Types

Database Port Types



Database JDBC Connection Port View



Database Connection Port View

Database Connector

Database Table Selector

Configure... F6

Execute F7

Execute and Open Views Shift+F10

Cancel F9

Reset F8

Edit Node Description... Alt+F2

New Workflow Annotation

Collapse into Meta Node

Compare Nodes

Show Flow Variable Ports

Cut

Copy

Paste

Undo

Redo

Delete

Database JDBC Connection

Database Connection - 2:42 - Database Table Selector(...)

File

Table Preview

Connection

Spec - Columns: 5

Flow Variables

Cache no. of rows: 100

Row ID	D sepal le...	D sepal w...	D petal le...	D petal wi...	S class
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1			
Row4	5	3.6			
Row5	5.4	3.9			
Row6	4.6	3.4			
Row7	5	3.4			
Row8	4.4	2.9			
Row9	4.9	3.1			
Row10	5.4	3.7			
Row11	4.8	3.4			
Row12	4.8	3			

Database Connection - 2:42 - Database Table Selector(...)

File

Table Preview

Connection

Spec - Columns: 5

Flow Variables

Database Driver: org.postgresql.Driver

Database URL:
jdbc:postgresql://192.168.56.111:5432/

User Name: knime

Database Type: postgresql

SQL Statement:
SELECT * FROM iris

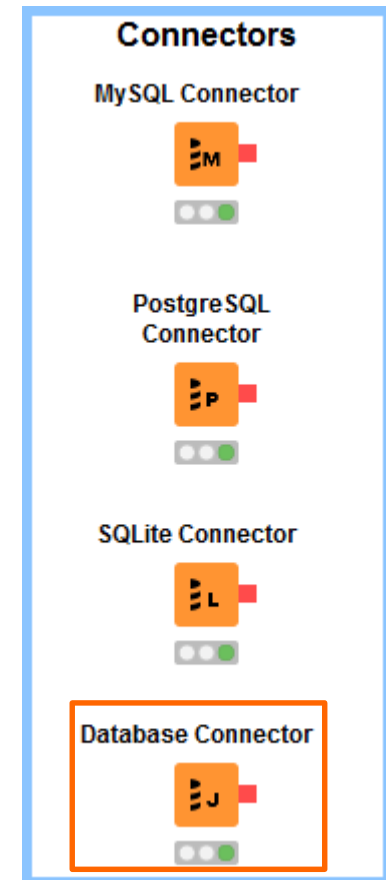
Copy SQL statement

Copy SQL to clipboard

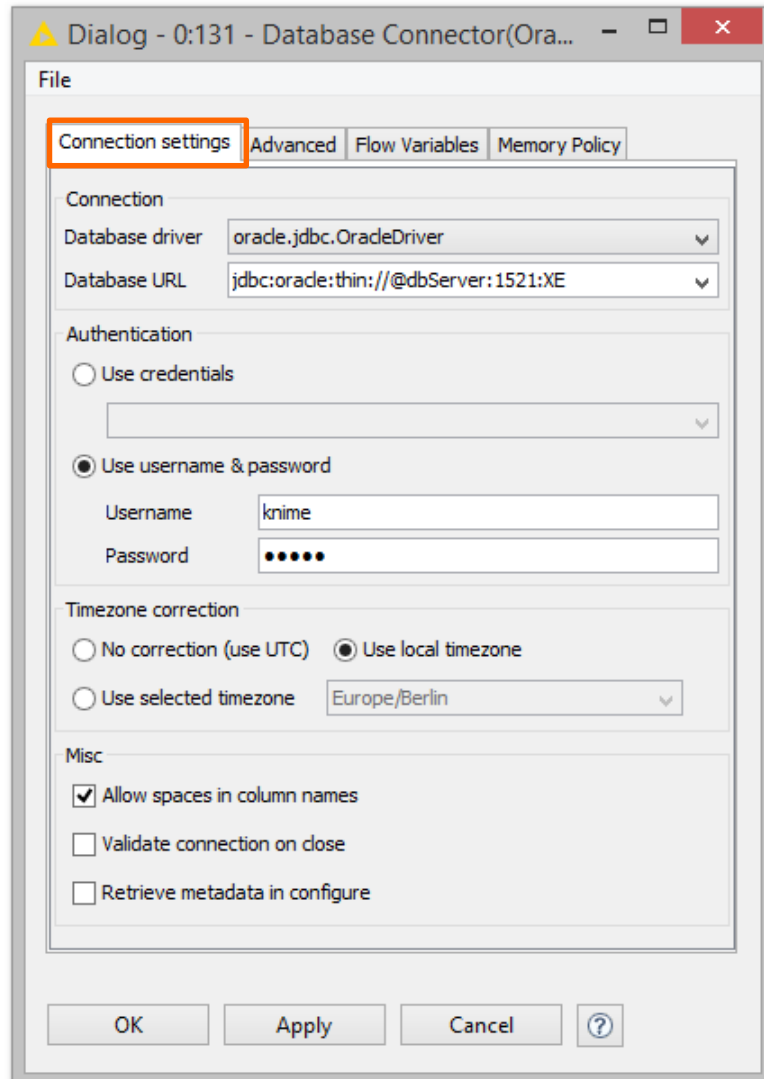
Connect to Database and Import Data

Database Connectors

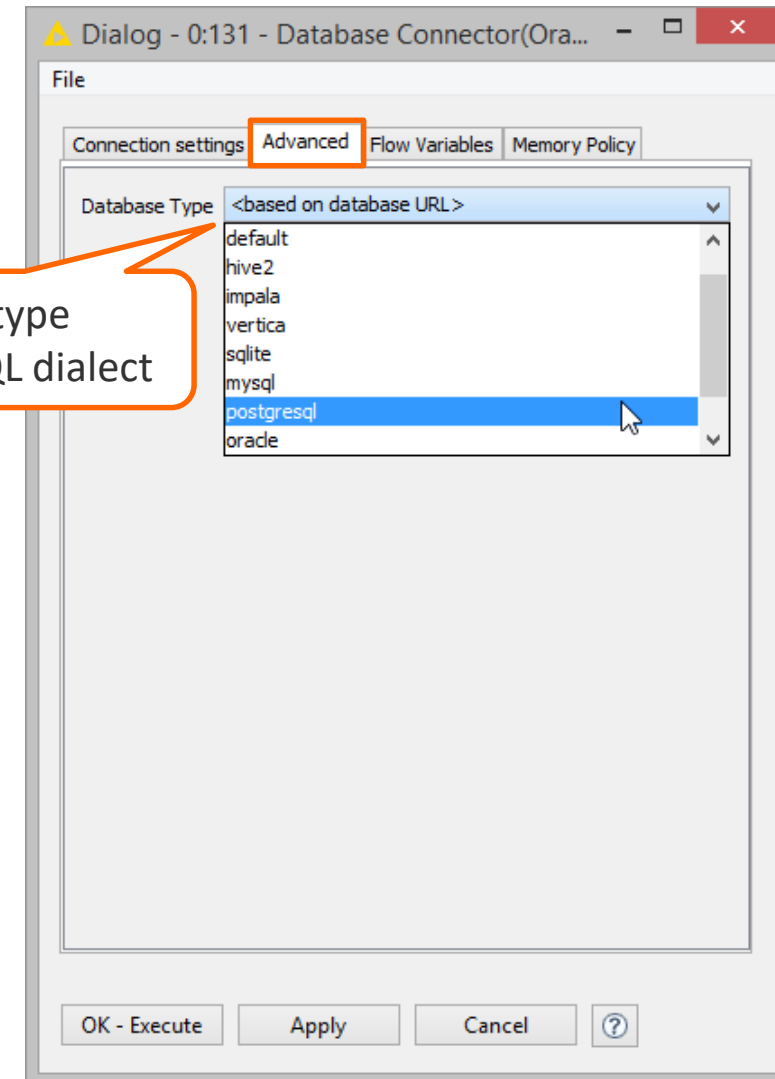
- Dedicated nodes to connect to specific Databases
 - Necessary JDBC driver included
 - Easy to use
 - Import DB specific behavior/capability
- Hive and Impala connector part of the commercial KNIME Big Data Connectors extension
- General Database Connector
 - Can connect to any JDBC source
 - Register new JDBC driver via
File -> Preferences -> KNIME -> Databases



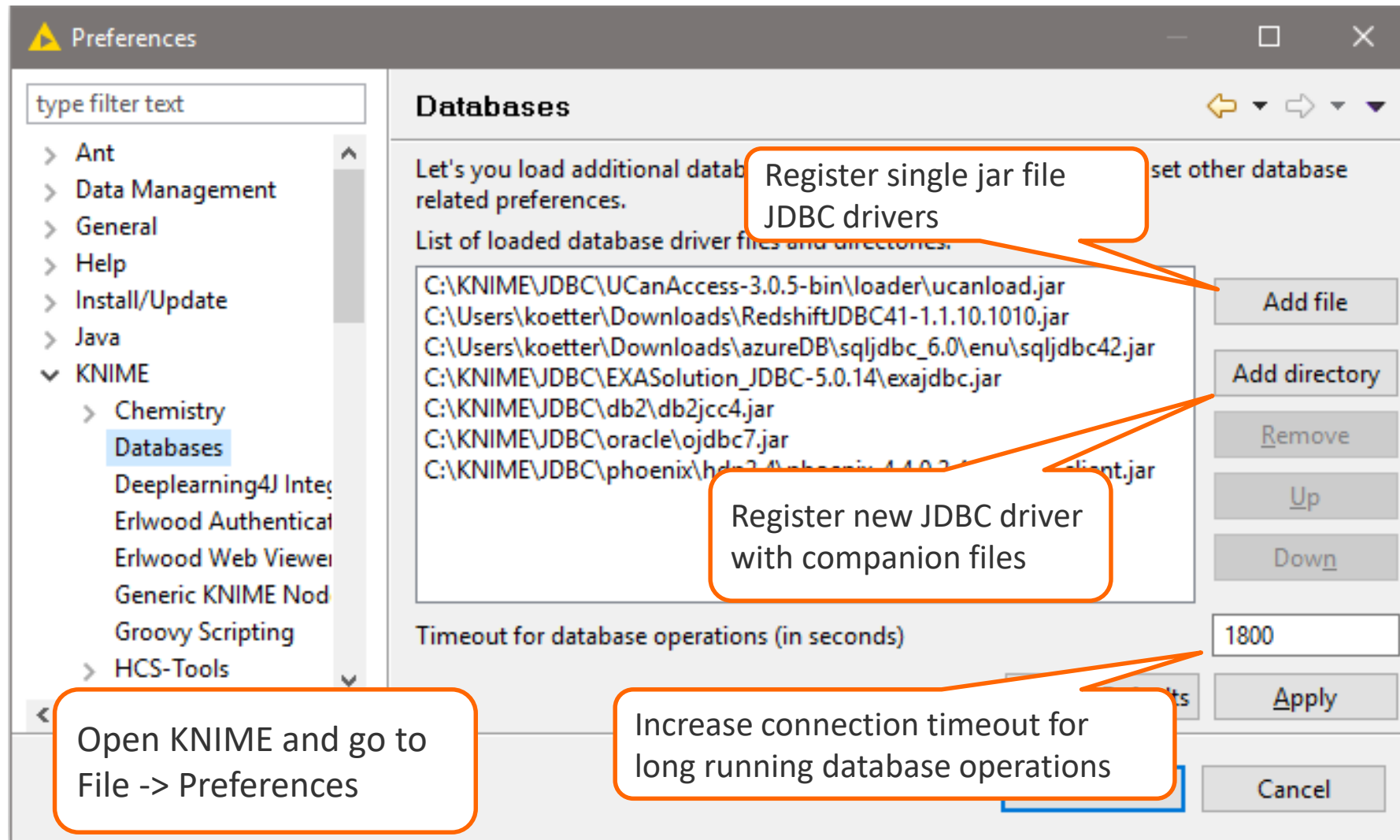
Database Connector node



Database type
defines SQL dialect

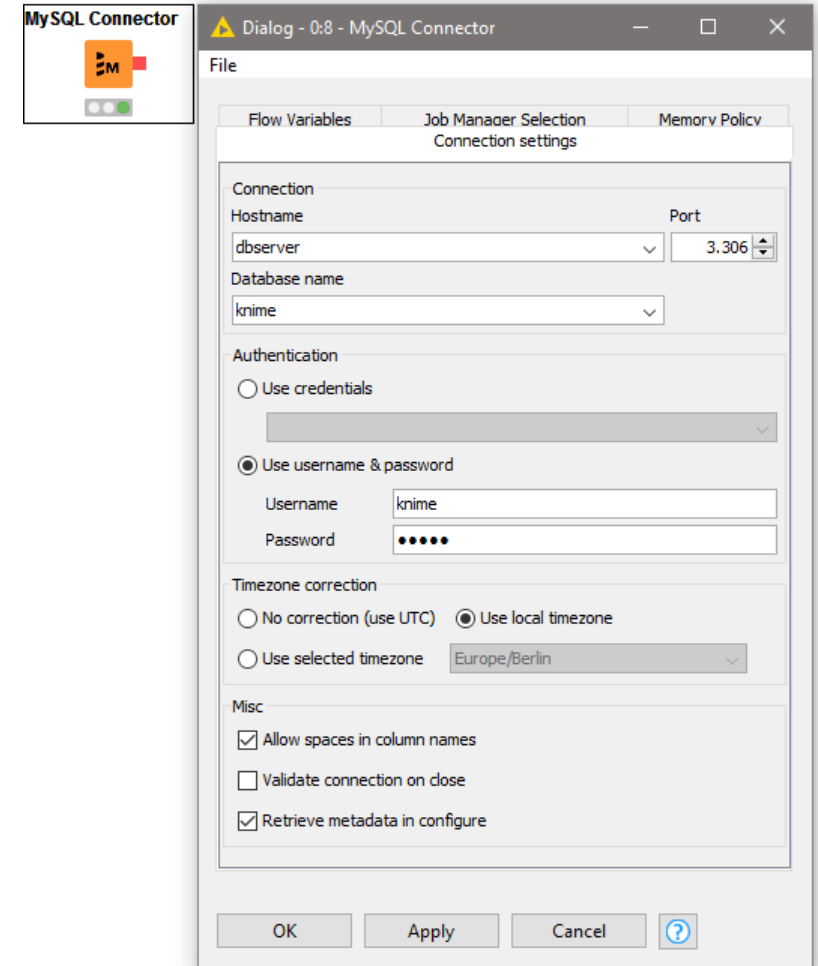
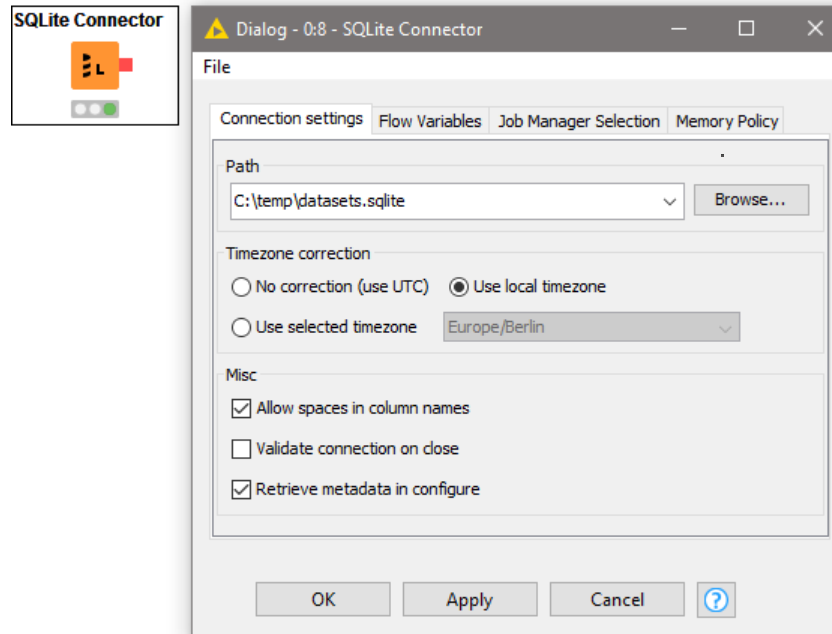
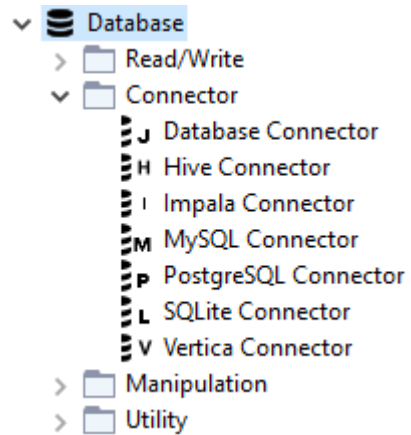


Register JDBC Driver



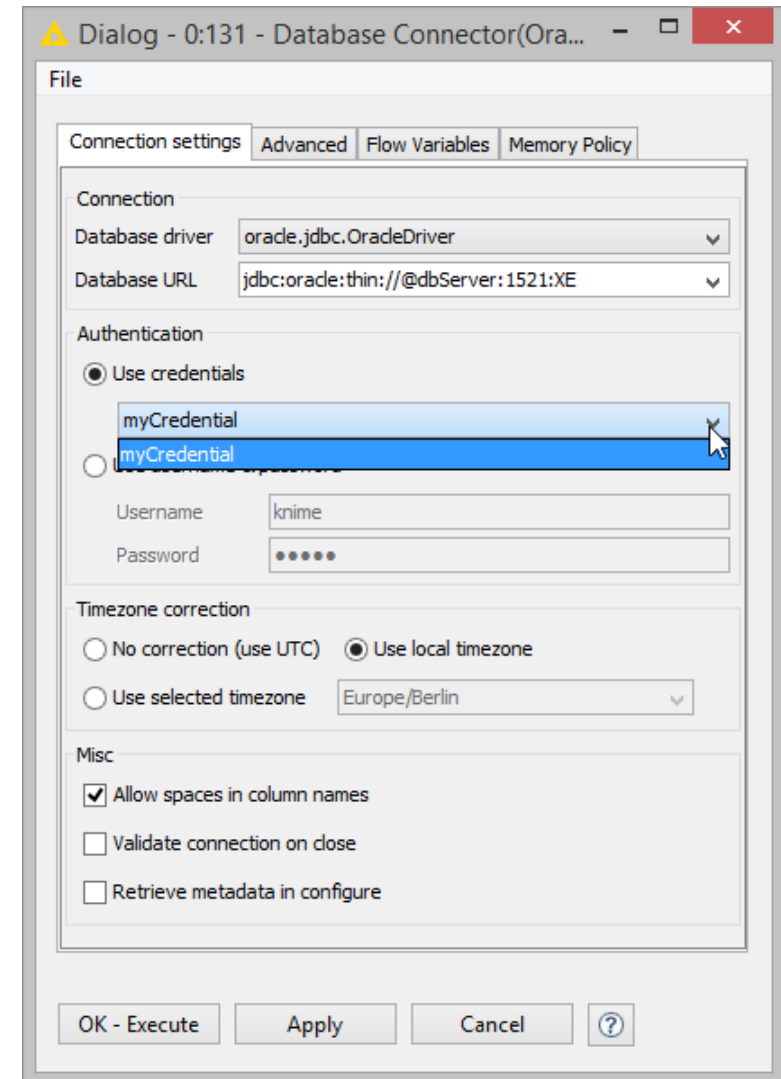
Dedicated Database Connectors

- MySQL, Postgres, SQLite and generic connectors.
- Propagate connection information to other DB nodes



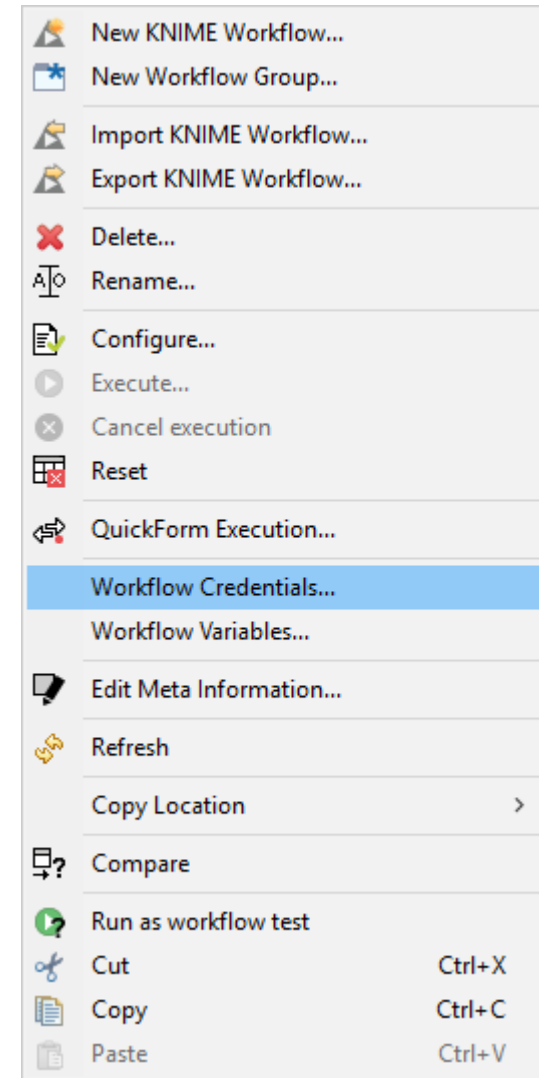
Workflow Credentials – Usage

- Replaces username and password fields
- Supported by several nodes that require login credentials
 - DB connectors
 - Remote file system connectors
 - Send mail
 - ...

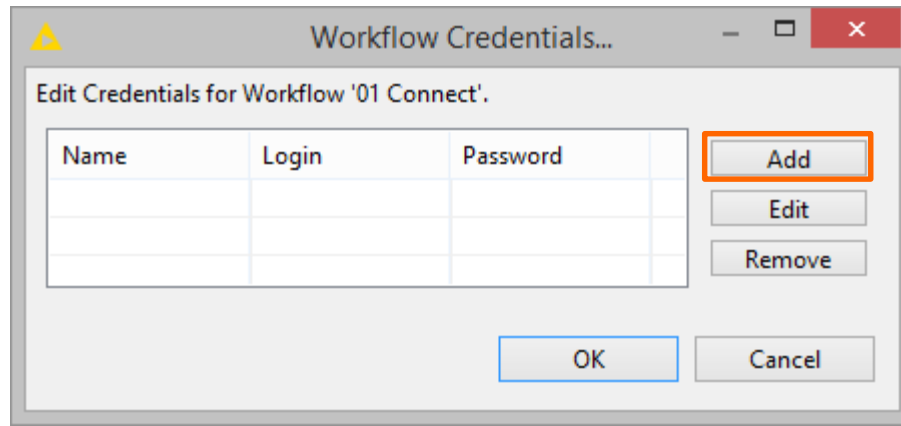


Workflow Credentials - Definition

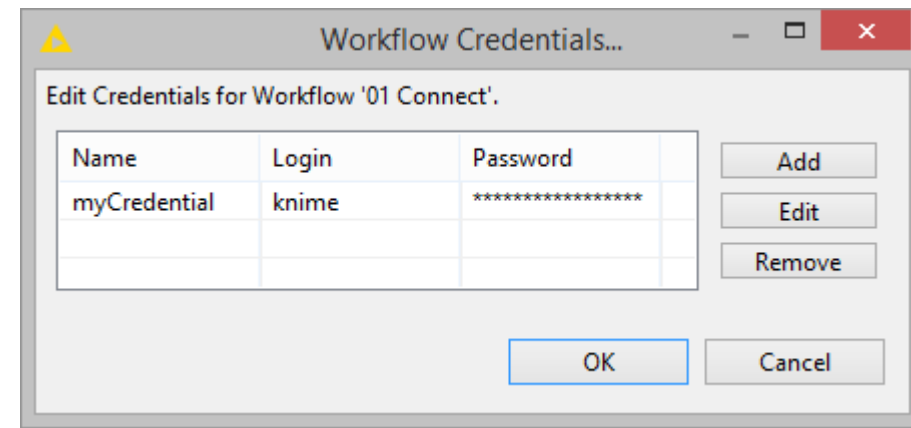
- Workflow needs to be open
- Right mouse click on workflow in KNIME explorer opens context menu
- Click on Workflow Credentials...



Workflow Credentials - Definition



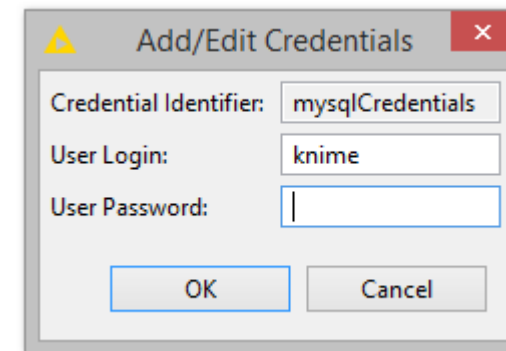
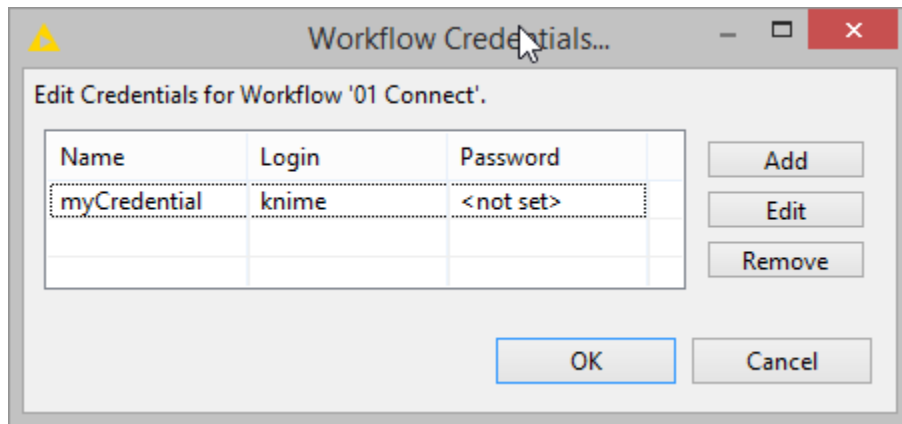
The 'Add/Edit Credentials' dialog box is shown. It has three input fields: 'Credential Identifier' with the value 'myCredential', 'User Login' with the value 'knime', and 'User Password' with masked characters '.....'. At the bottom are 'OK' and 'Cancel' buttons.



- You can define multiple credentials for different databases

Workflow Credentials – Open Workflow with Credentials

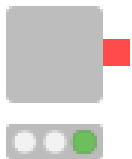
- Shows Workflow Credentials when workflow is opened
- Double click on entry to set password



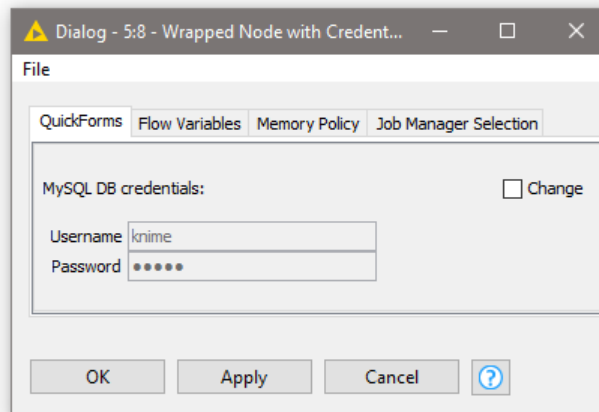
Credentials Input Quickform Node

- Will replace workflow credentials
- Works together with all nodes that support workflow credentials

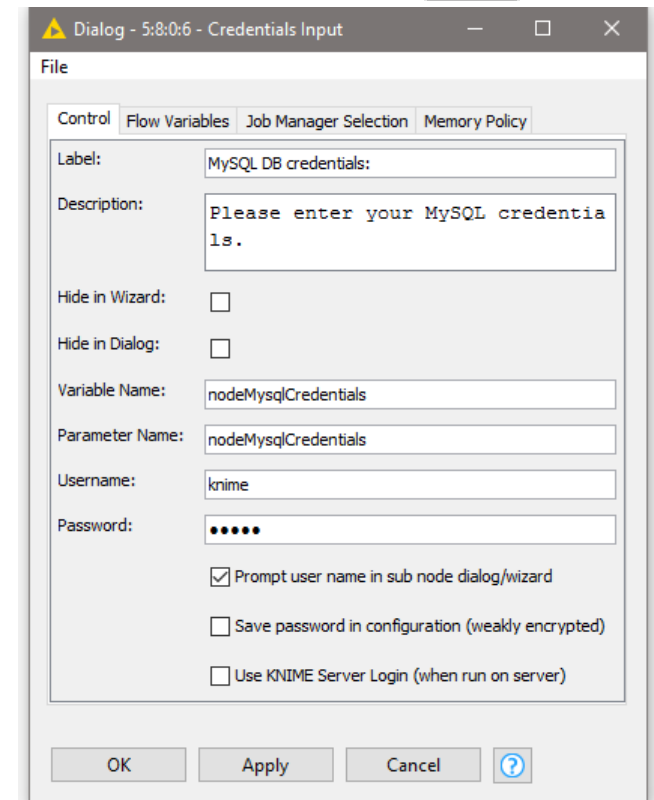
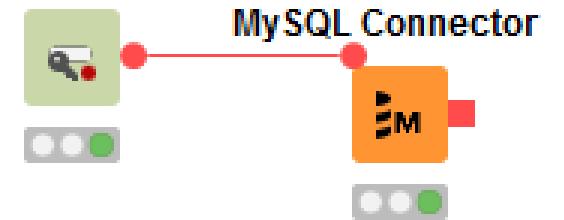
Wrapped Node with Credentials Input



Ctrl+ double click
to look inside

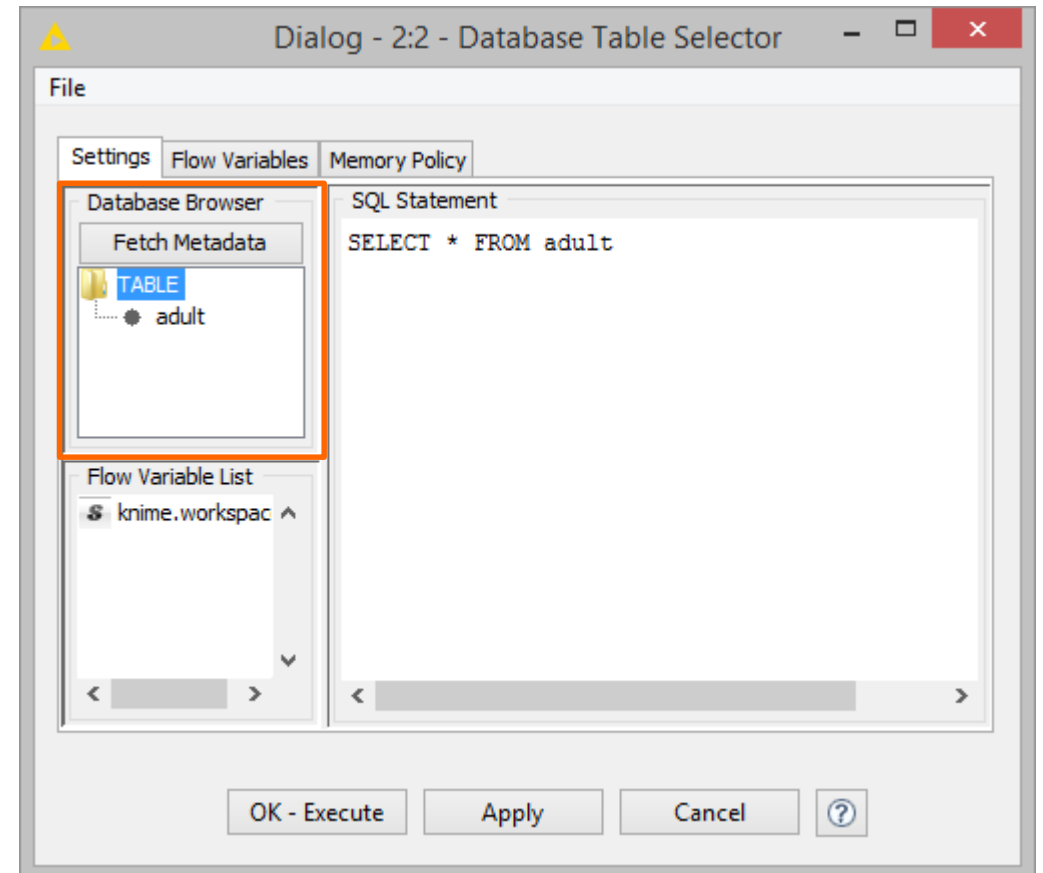
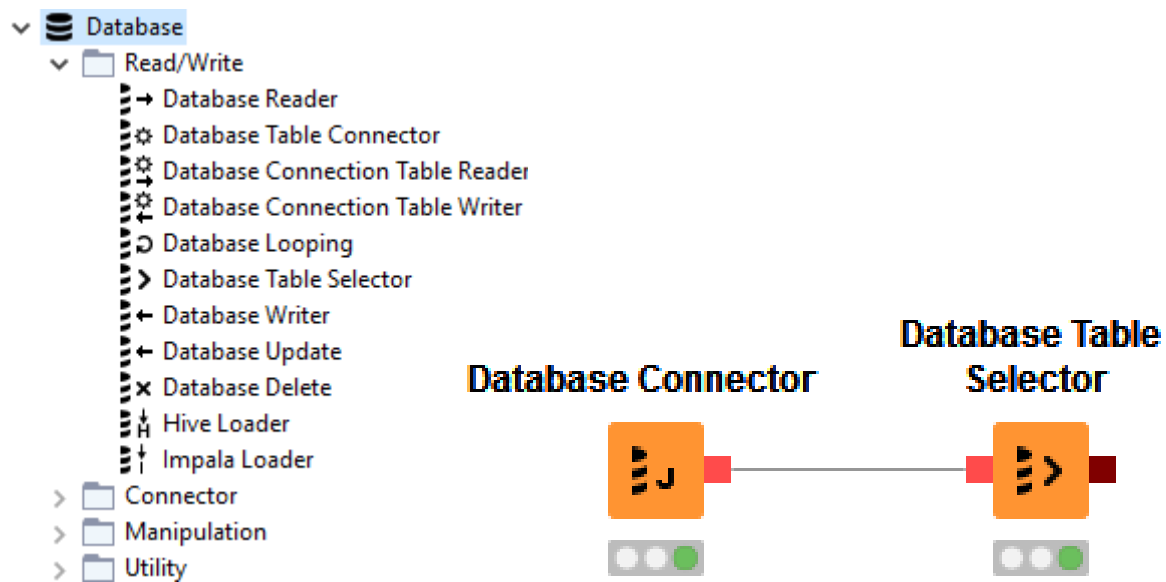


Credentials Input



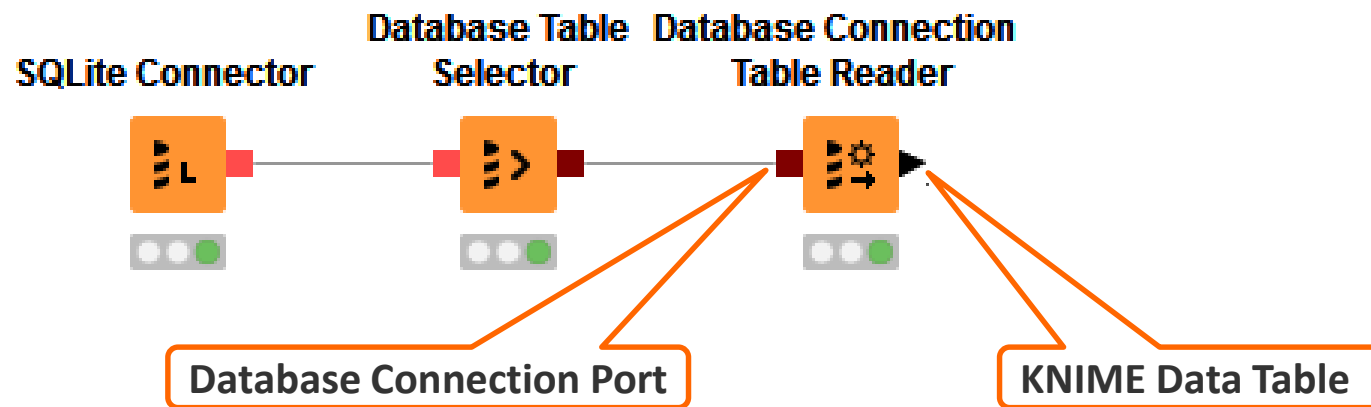
Database Table Selector

- Takes connection information and constructs a query
- Explore DB metadata
- Outputs a SQL query



Database Connection Table Reader

- Executes incoming SQL Query on Database
- Reads results into a KNIME data table



Section Exercise – 01_DB_Connect

- Connect to the database (SQLite) newCensus.sqlite in folder 1_Data
 - Use SQLite Connector (Note: SQLite Connector supports knime:// protocol)
- Explore DB metadata
- Select table ss13pme (person data in Maine)
- Import the data into a KNIME data table

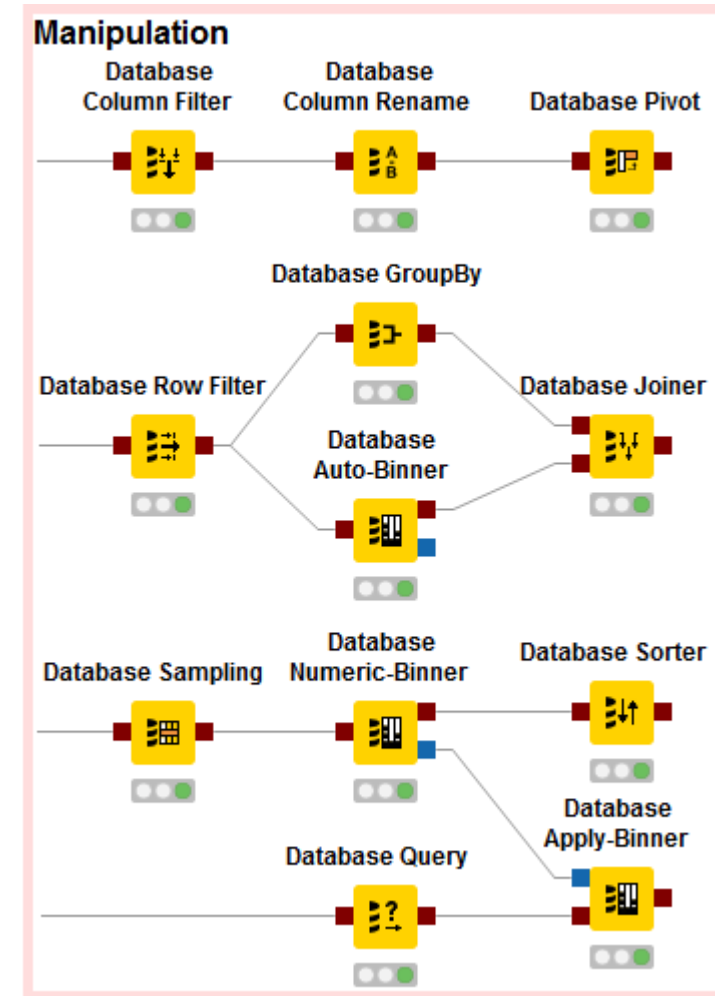
Optional: Create a workflow credential and use it in a MySQL Connector instead of user name and password. Create a Credentials Input node and use it in another MySQL Connector instead of user name and password.

In-Database Processing



Query Nodes

- Filter rows and columns
- Join tables/queries
- Extract samples
- Bin numeric columns
- Sort your data
- Write your own query
- Aggregate your data



Data Aggregation

Rowid	Group	Value
r1	M	2
r2	F	3
r3	M	1
r4	F	5
r5	F	7
r6	M	5

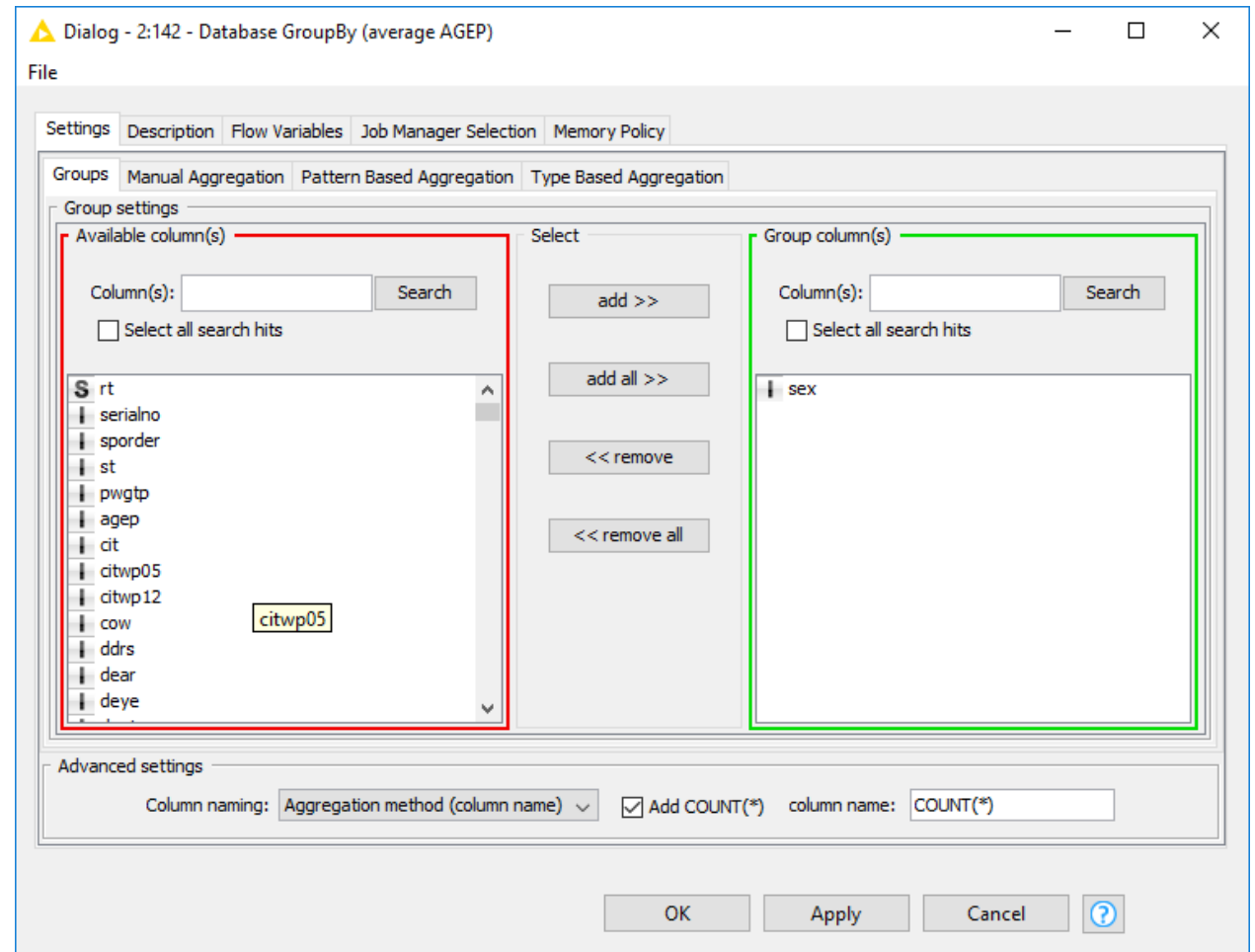
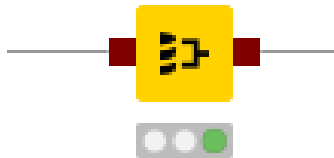
Rowid	Group	Value
r1+r3+r6	M	8
r2+r4+r5	F	15

aggregated on “Group” by method: sum(“Value”)

Database GroupBy

- Aggregate to summarize data

Database GroupBy



Database GroupBy – Manual Aggregation

Dialog - 2:142 - Database GroupBy (average AGEP)

File

SettingsDescriptionFlow VariablesJob Manager SelectionMemory Policy

GroupsManual AggregationPattern Based AggregationType Based Aggregation

Aggregation settings

Available columns

rt
serialno
sporder
st
pwgtp
agep
cit
citwp05
citwp12
cow
ddrs
dear
deye
dout
dphy
drat
dratx
drem

Select

add >>
add all >>
<< remove
<< remove all

To change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Parameter
agep	AVG	Edit
agep	AVG	Edit
	COUNT	
	GROUP_CONCAT	
	MAX	
	MIN	
	SUM	
	TOTAL	
	custom	

Advanced settings

Column naming: Aggregation method (column name) Add COUNT(*) column name: COUNT(*)

OKApplyCancel?

Returns number of rows per group

Copyright © 2017 KNIME.com AG

74

Licensed under a Creative Commons Attribution-Noncommercial-Share Alike license
<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Open for Innovation
KNIME

Database GroupBy – Pattern Based Aggregation

The screenshot shows the 'Pattern Based Aggregation' tab of the 'Database GroupBy' dialog. The 'Aggregation settings' section contains a table with the following data:

Search pattern (double click to change)	Regex	Aggregation (click to change)	Parameter
*	☒	custom	Edit

A callout box points to the 'Regex' checkbox with the text: "Tick this option if the search pattern is a regular expression otherwise it is treated as string with wildcards ('*' and '?')".

The 'Advanced settings' section at the bottom shows 'Column naming' set to 'Aggregation method (column name)', 'Add COUNT(*)' checked, and 'column name' set to 'COUNT(*)'.

Database GroupBy – Type Based Aggregation

Matches all columns

Matches all numeric columns

Settings | Description | Flow Variables | Memory Policy

Groups | Manual Aggregation | Pattern Based Aggregation | **Type Based Aggregation**

Aggregation settings

Available data type

- B BooleanCell
- ? DataCell
- 📅 DateAndTimeCell
- D DoubleCell**
- I IntCell
- (...) ListCell (Collection of:)
- L LongCell
- (...) SetCell (Collection of:)
- S StringCell

Select

add >>

add all >>

<< remove

<< remove all

To change multiple columns use right mouse click for context menu.

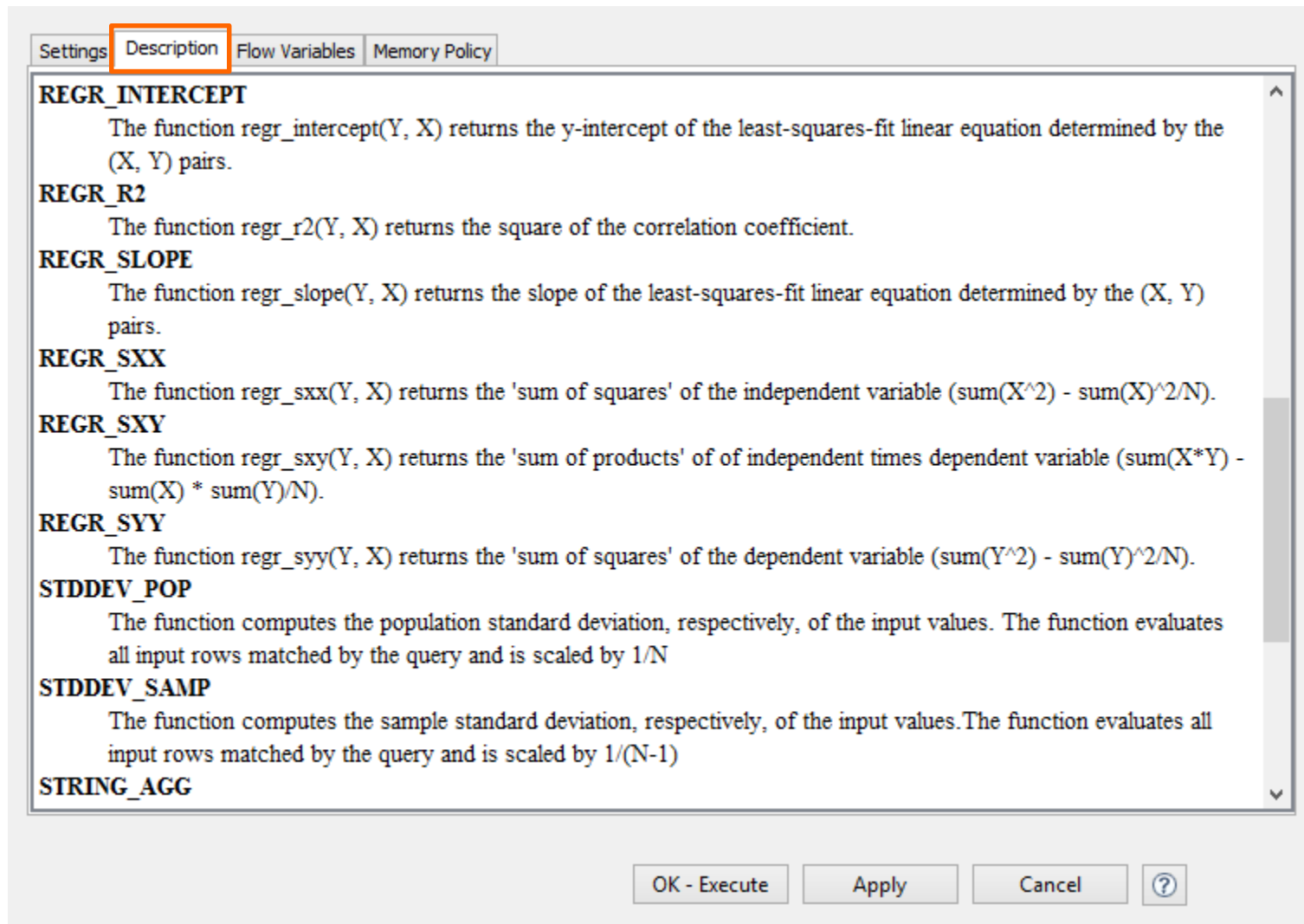
Data type	Aggregation (click to change)	Parameter
D DoubleCell	AVG	
D DoubleCell	SUM	

Advanced settings

Column naming: Aggregation method (column name) v ☒ Add COUNT(*) column name: COUNT(*)

OK - Execute Apply Cancel ?

Database GroupBy – Aggregation Method Description



Settings **Description** Flow Variables Memory Policy

REGR_INTERCEPT
The function `regr_intercept(Y, X)` returns the y-intercept of the least-squares-fit linear equation determined by the (X, Y) pairs.

REGR_R2
The function `regr_r2(Y, X)` returns the square of the correlation coefficient.

REGR_SLOPE
The function `regr_slope(Y, X)` returns the slope of the least-squares-fit linear equation determined by the (X, Y) pairs.

REGR_SXX
The function `regr_sxx(Y, X)` returns the 'sum of squares' of the independent variable ($\sum(X^2) - \sum(X)^2/N$).

REGR_SXY
The function `regr_sxy(Y, X)` returns the 'sum of products' of independent times dependent variable ($\sum(X*Y) - \sum(X) * \sum(Y)/N$).

REGR_SYY
The function `regr_syy(Y, X)` returns the 'sum of squares' of the dependent variable ($\sum(Y^2) - \sum(Y)^2/N$).

STDDEV_POP
The function computes the population standard deviation, respectively, of the input values. The function evaluates all input rows matched by the query and is scaled by $1/N$.

STDDEV_SAMP
The function computes the sample standard deviation, respectively, of the input values. The function evaluates all input rows matched by the query and is scaled by $1/(N-1)$.

STRING_AGG

OK - Execute Apply Cancel ?

Database GroupBy – DB Specific Aggregation Methods

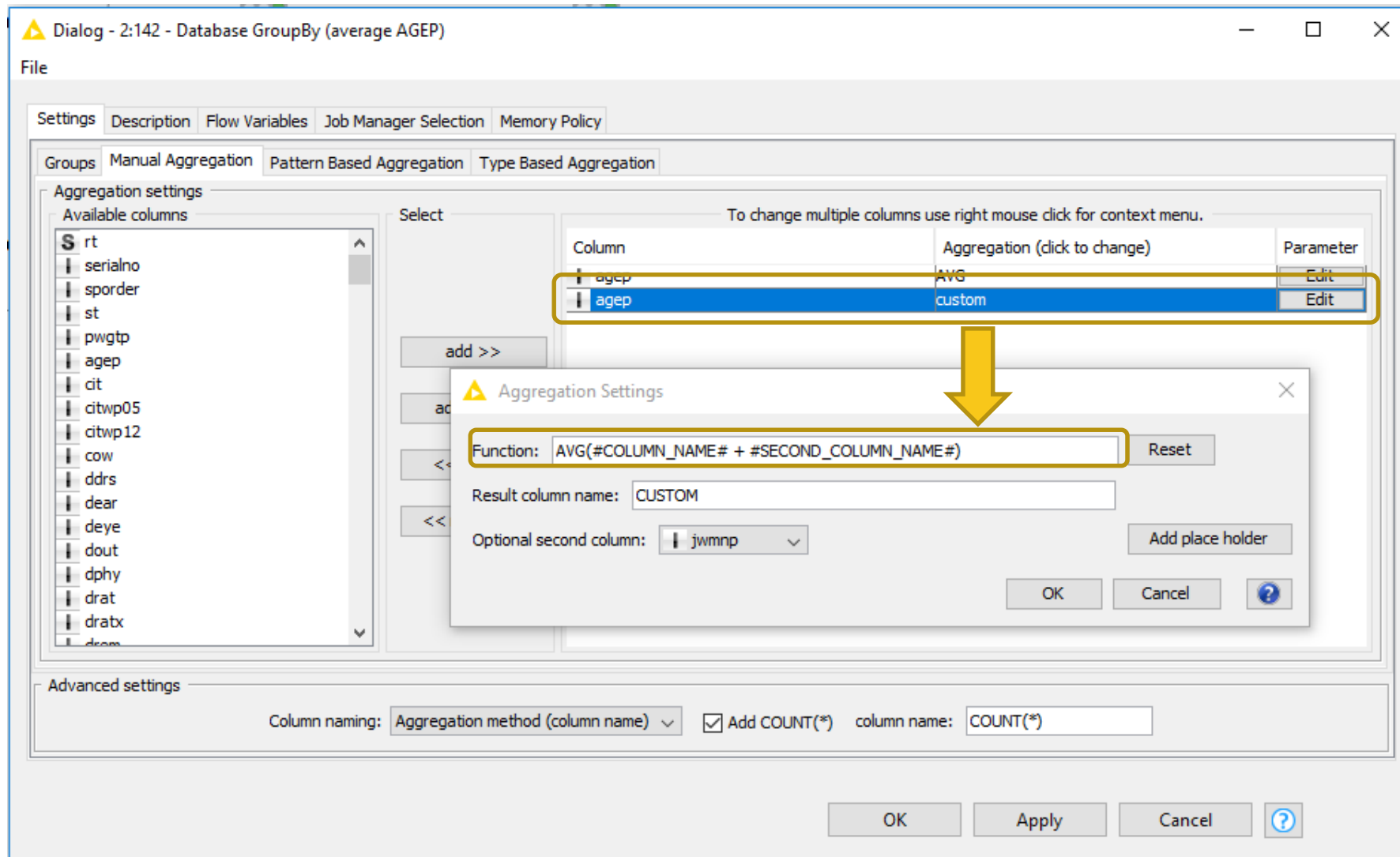
Column	Aggregation (click to change)	Parameter
D sepal length	AVG	
D sepal width	AVG	Edit
D petal length	COUNT	Edit
D petal width	GROUP_CONCAT	
S id	MAX	
D sepal width	MIN	
D petal length	SUM	
S id	TOTAL	Edit
S id	custom	Edit
D petal length	GROUP_CONCAT	Edit

SQLite: 7 aggregation functions

Column	Aggregation (click to change)	Parameter
D sepal length	ARRAY_AGG	
D sepal width	ARRAY_AGG	
D petal length	AVG	Edit
D petal width	CORR	Edit
D sepal length	COUNT	Edit
D sepal width	COVAR_POP	Edit
D petal length	COVAR_SAMP	Edit
D petal width	MAX	
D sepal length	MIN	
D sepal width	REGR_AVGX	Edit
D petal length	REGR_AVGY	Edit
D petal width	REGR_COUNT	Edit
D sepal length	REGR_INTERCEPT	Edit
D sepal width	REGR_R2	Edit
D petal length	REGR_SLOPE	Edit
D petal width	REGR_SXX	Edit
D sepal length	REGR_SXY	Edit

PostgreSQL: 25 aggregation functions

Database GroupBy – Custom Aggregation Function



Joining Columns of Data

Data read from Database - 2:3 - Database C...

File

Table "database" - Rows: 4 Spec - Columns: 4 Properties Flow Variables

Row ID	id	age	sex	income
Row0	1	41	Male	<=50K
Row1	3	19	Male	<=50K
Row2	6	34	Female	<=50K
Row3	8	22	Male	<=50K

Left Table

Join by id

Inner Join

Data read from Database - 2:9 - Database Co...

File

Table "database" - Rows: 6 Spec - Columns: 4 Properties Flow Variables

Row ID	id	age	occupation	income
Row0	1	41	Other-service	<=50K
Row1	2	44	Sales	<=50K
Row2	4	64	Transport-m...	<=50K
Row3	5	50	Craft-repair	>50K
Row4	6	34	Other-service	<=50K
Row5	7	38	Other-service	<=50K

Right Table

Output data - 2:12 - Column Resorter

File

Table "default" - Rows: 2 Spec - Columns: 8 Properties Flow Variables

Row ID	id	age	income	sex	id (#1)	age (#1)	income (#1)	occupation
Row0	1	41	<=50K	Male	1	41	<=50K	Other-service
Row1	6	34	<=50K	Female	6	34	<=50K	Other-service

Right Outer Join

Left Outer Join

Output data - 2:15 - Column Resorter

File

Table "default" - Rows: 4 Spec - Columns: 8 Properties Flow Variables

Row ID	id	age	income	sex	id (#1)	age (#1)	income (#1)	occupation
Row0	1	41	<=50K	Male	1	41	<=50K	Other-service
Row2	6	34	<=50K	Female	6	34	<=50K	Other-service
Row1	3	19	<=50K	Male	?	?	?	?
Row3	8	22	<=50K	Male	?	?	?	?

Missing values in the right table.

Output data - 2:21 - Column Resorter

File

Table "default" - Rows: 6 Spec - Columns: 9 Properties Flow Variables

Row ID	id	age	income	sex	id (#1)	age (#1)	income (#1)	occupation
Row0	1	41	<=50K	Male	1	41	<=50K	Other-service
Row4	6	34	<=50K	Female	6	34	<=50K	Other-service
Row1	?	?	?	?	2	44	<=50K	Sales
Row2	?	?	?	?	4	64	<=50K	Transport-moving
Row3	?	?	?	?	5	50	>50K	Craft-repair
Row5	?	?	?	?	7	38	<=50K	Other-service

Missing values in the left table.

Joining Columns of Data

Data read from Database - 2:3 - Database C...

File

Table "database" - Rows: 4 Spec - Columns: 4 Properties Flow Variables

Row ID	id	age	\$ sex	\$ income
Row0	1	41	Male	<=50K
Row1	3	19	Male	<=50K
Row2	6	34	Female	<=50K
Row3	8	22	Male	<=50K

Left Table

Join by id

Full Outer Join

Data read from Database - 2:9 - Database Co...

File

Table "database" - Rows: 6 Spec - Columns: 4 Properties Flow Variables

Row ID	id	age	\$ occupa...	\$ income
Row0	1	41	Other-service	<=50K
Row1	2	44	Sales	<=50K
Row2	4	64	Transport-m...	<=50K
Row3	5	50	Craft-repair	>50K
Row4	6	34	Other-service	<=50K
Row5	7	38	Other-service	<=50K

Right Table

Output data - 2:29 - Column Resorter

File

Table "default" - Rows: 8 Spec - Columns: 8 Properties Flow Variables

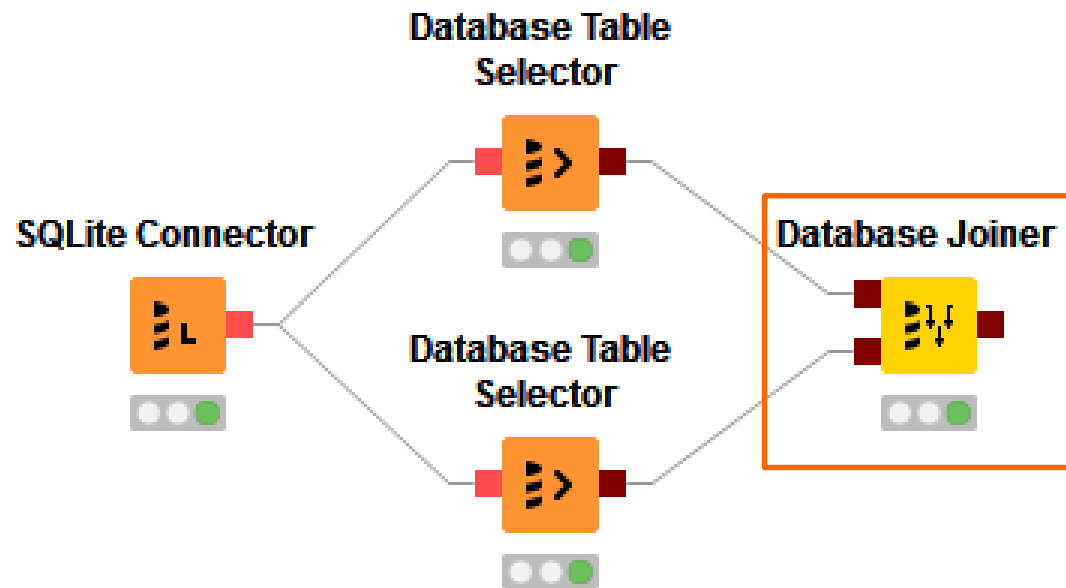
Row ID	id	age	\$ income	\$ sex	id (#1)	age (#1)	\$ income (#1)	\$ occupation
Row0	1	41	<=50K	Male	1	41	<=50K	Other-service
Row7	3	19	<=50K	Male	?	?	?	?
Row4	6	34	<=50K	Female	6	34	<=50K	Other-service
Row6	8	22	<=50K	Male	?	?	?	?
Row1	?	?	?	?	2	44	<=50K	Sales
Row2	?	?	?	?	4	64	<=50K	Transport-moving
Row3	?	?	?	?	5	50	>50K	Craft-repair
Row5	?	?	?	?	7	38	<=50K	Other-service

Missing values in the left table.

Missing values in the right table.

Database Joiner

- Combines columns from 2 different tables
- Top port contains “Left” data table
- Bottom port contains the “Right” data table



Joiner Configuration – Linking Rows

Dialog - 2:177 - Database Joiner (inner join)

File

Joiner Settings | Column Selection | Flow Variables | Job Manager Selection | Memory Policy

Join Mode

Join mode: Inner Join

Joining Columns

☒ Match all of the following ☐ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)	
SERIALNO	serialno	+
		+

OK Apply Cancel ?

Values to join on.
Multiple joining
columns are allowed.

Joiner Configuration – Column Selection

Dialog - 2:177 - Database Joiner (inner join)

File

Joiner Settings **Column Selection** Flow Variables Job Manager Selection Memory Policy

Top Input (left' table)

Exclude

Column(s): Search

☐ Select all search hits

RT
PUMA00
PUMA10
ST
ADJHSG
ADJINC
WGTP1
WGTP2

Select

add >>

add all >>

<< remove

<< remove all

Include

Column(s): Search

☐ Select all search hits

insp
SERIALNO
DIVISION
REGION
WGTP
NP
TYPE
ACR

☐ Always include all columns

Bottom Input (right' table)

Exclude

Column(s): Search

☐ Select all search hits

Select

add >>

add all >>

<< remove

<< remove all

Include

Column(s): Search

☐ Select all search hits

rt
serialno
sporder
st
pwgtp
agep
cit
citwp05

☒ Always include all columns

Duplicate Column Handling

☐ Filter duplicates

☐ Don't execute

☒ Append suffix (automatic)

☐ Append custom suffix:

Joining Columns Handling

☐ Remove joining columns from top input ('left' table)

☒ Remove joining columns from bottom input ('right' table)

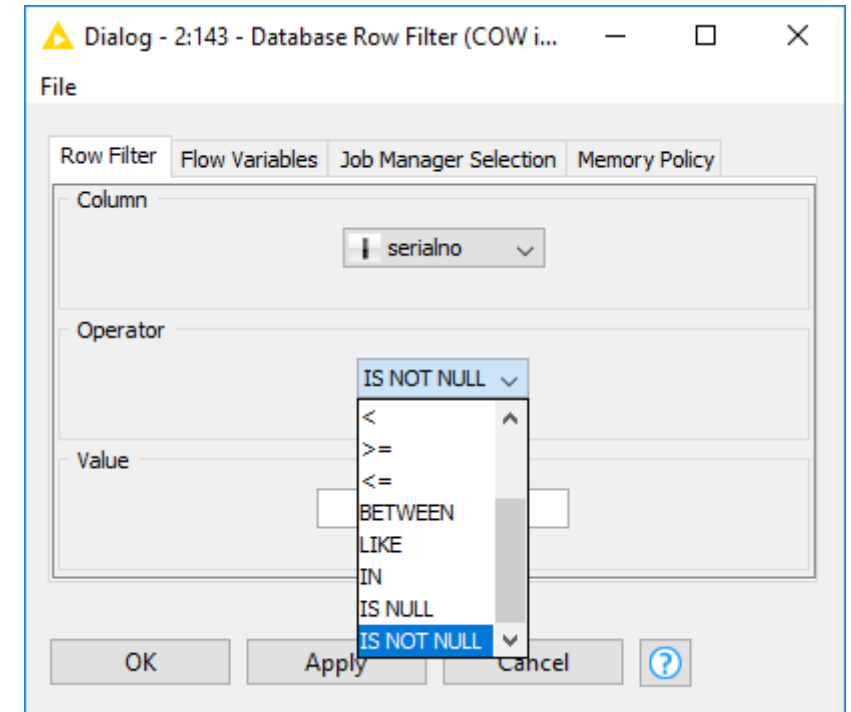
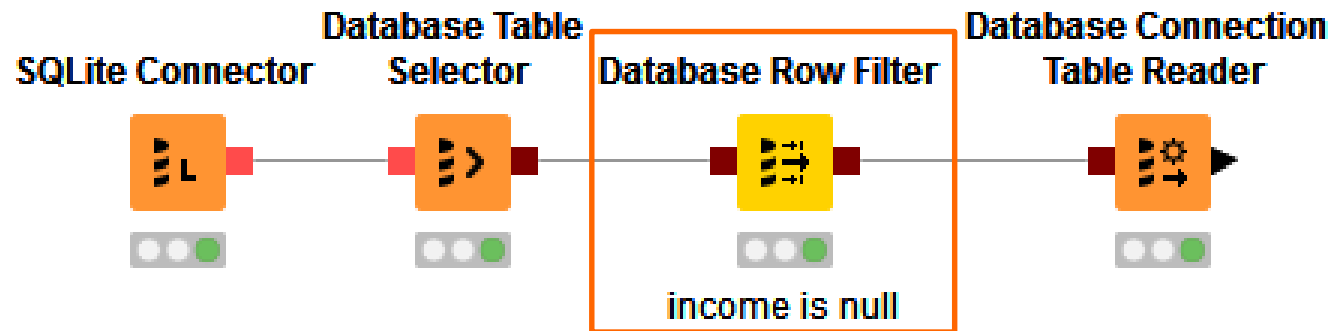
OK Apply Cancel ?

Columns from left table to output table

Columns from right table to output table

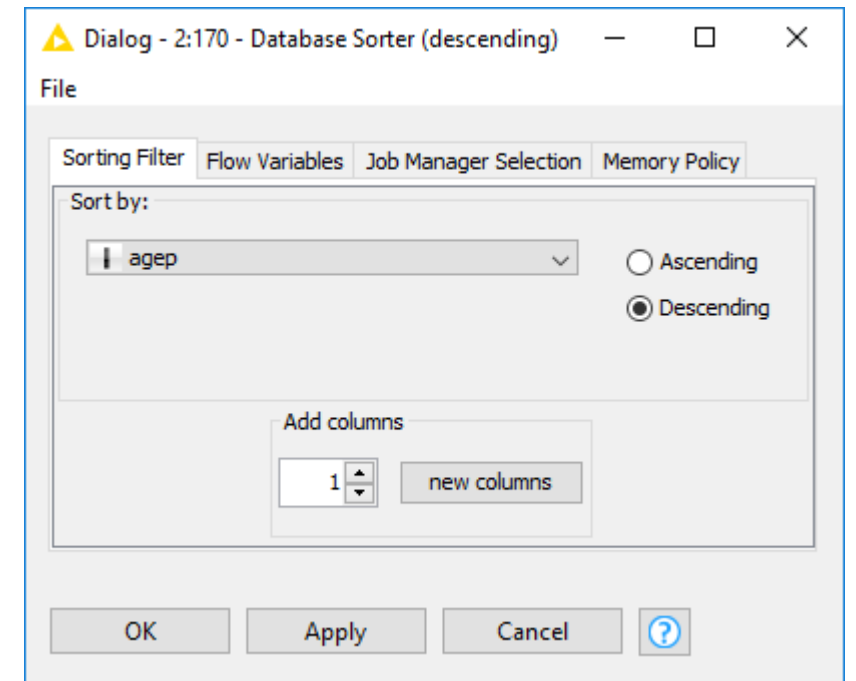
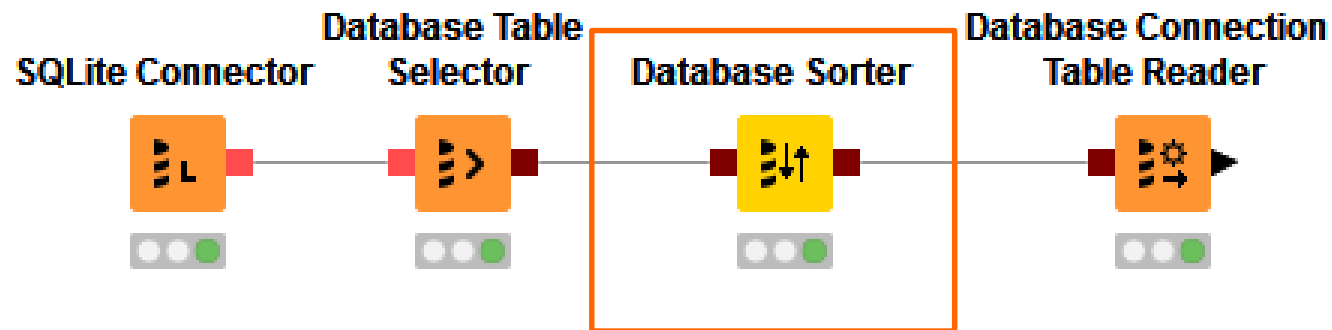
Database Row Filter

- Filters rows that do not match the filter criteria
- Use the *IS NULL* or *IS NOT NULL* operator to filter missing values



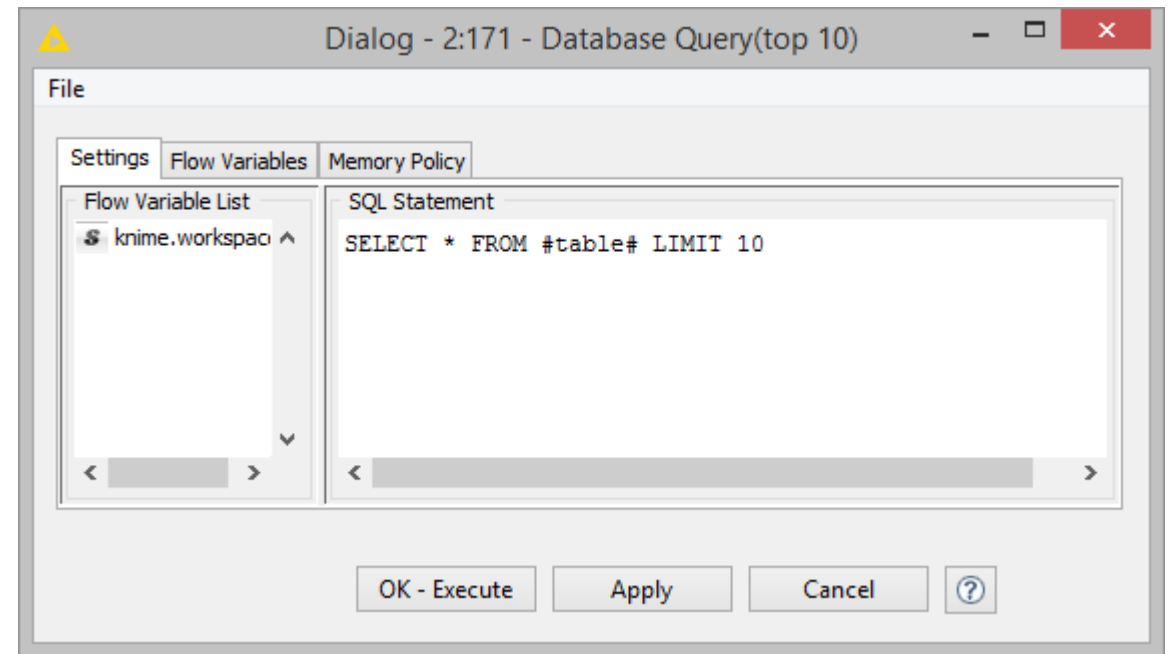
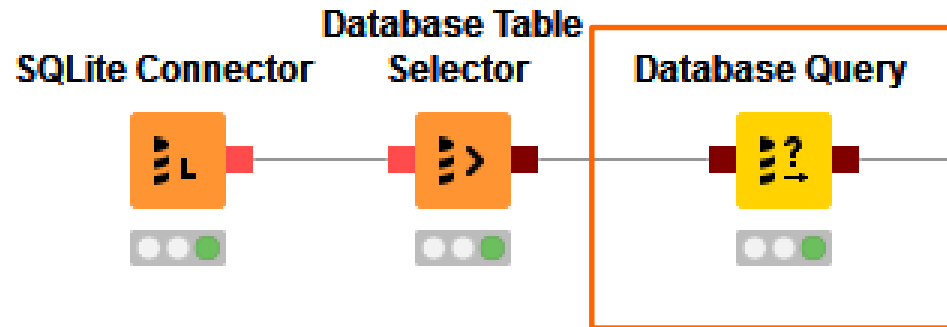
Database Sorter

- Sorts the input data by one or multiple columns



Database Query

- Executes arbitrary SQL queries
- #table# is replaced with input query



Section Exercise – 02_DB_InDB_Processing

From tables ss13hme (house data) and ss13pme (person data) in database newCensus.sqlite

- join ss13hme and ss13pme on SERIALNO. Remove all columns named PUMA* and PWGTP* from both tables.
- filters all rows from ss13pme where COW is NULL and where COW is NOT NULL
- calculate average AGEp for the different SEX groups
- For all tasks, at the end load data into KNIME

Optional. Sort the data rows by descending AGEp and extract top 10 only.

Hint: Use LIMIT to restrict the number of rows returned by the db.

Predicting income values with KNIME

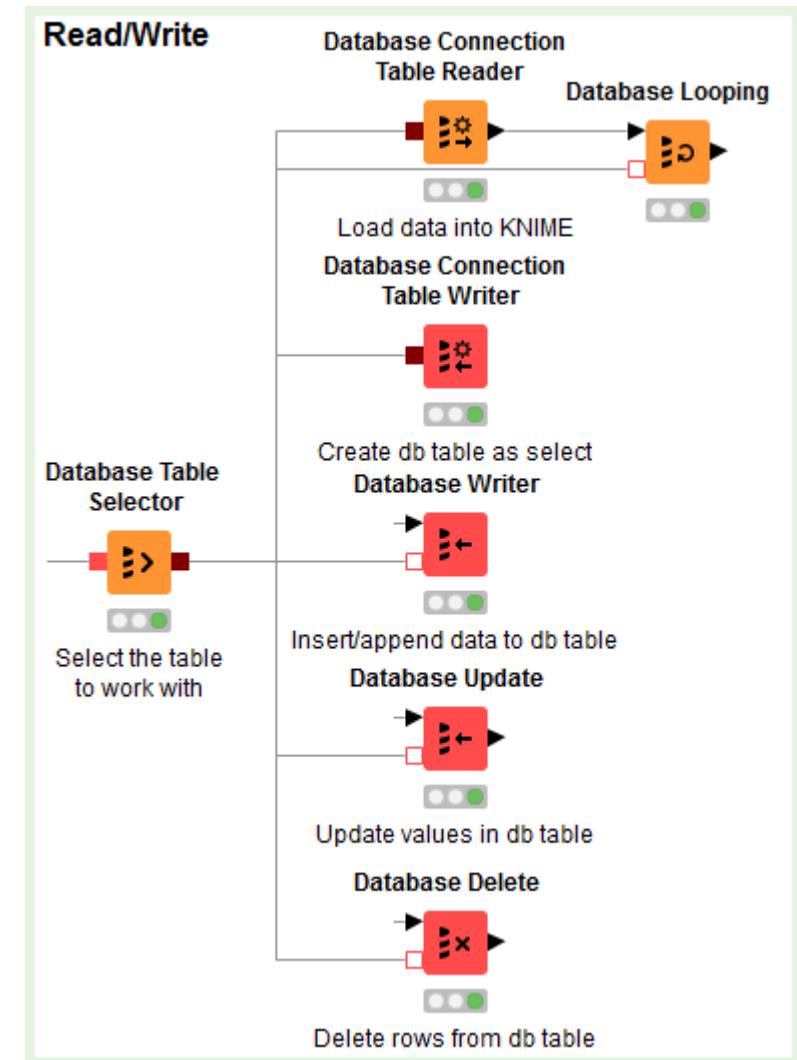
Section Exercise – 03_DB_Modelling

- Train a Decision Tree to predict the income where COW is not null
- Apply Decision Tree Model to predict income where COW is missing (null)

Write/Load Data into a Database

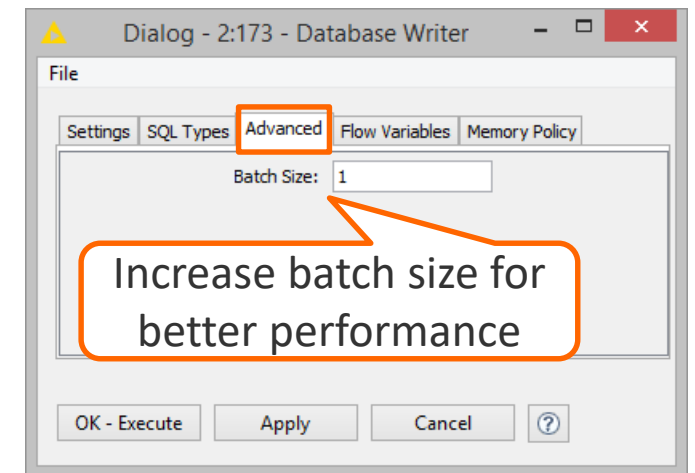
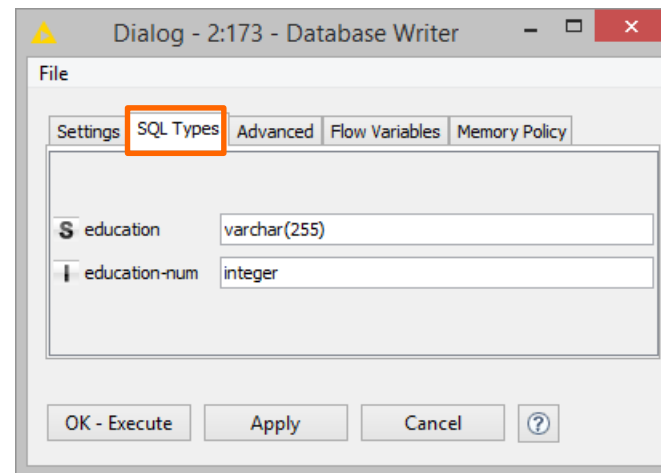
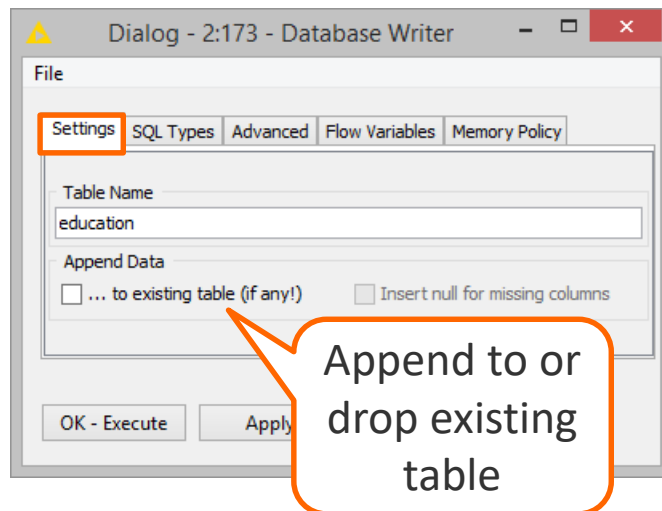
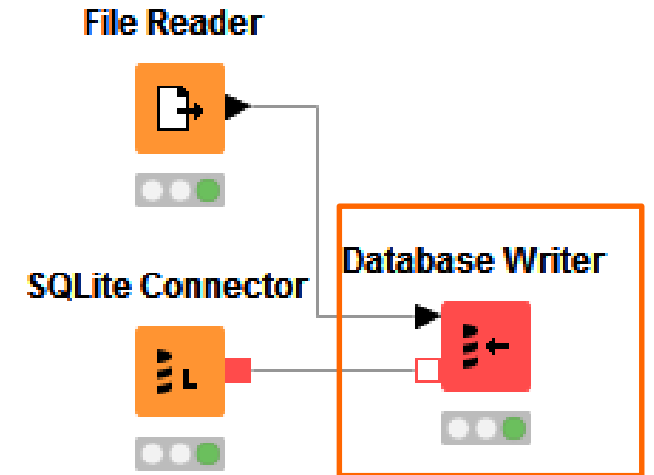
Database Writing Nodes

- Create table as select
- Insert/append data
- Update values in table
- Delete rows from table



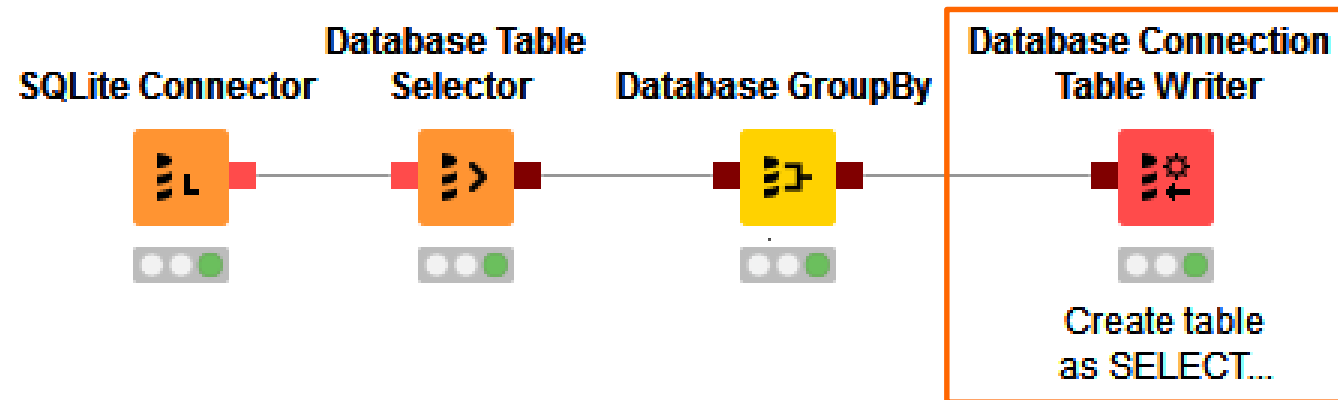
Database Writer

- Writes data from a KNIME data table **directly** into a database table



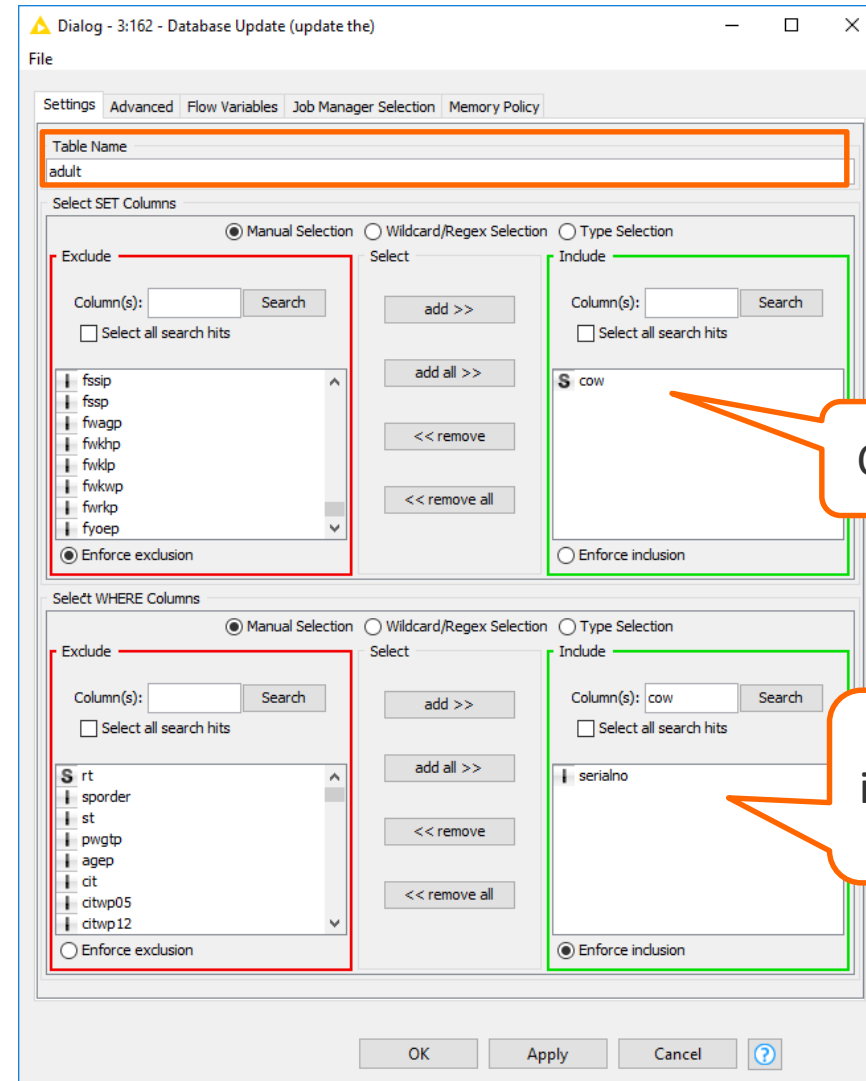
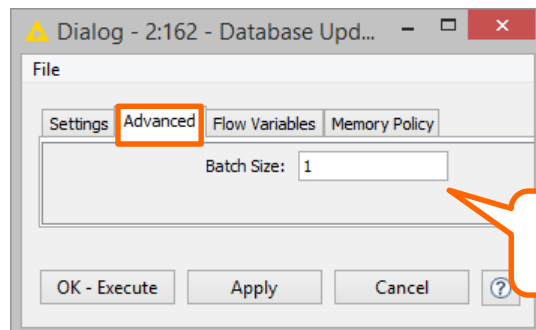
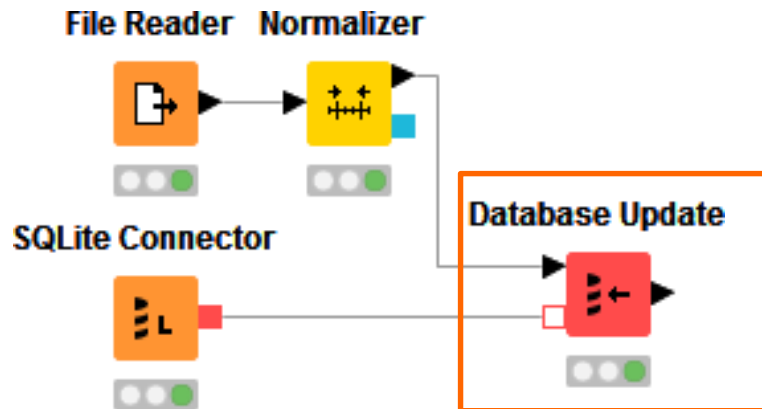
Database Connection Table Writer

- Creates a new database table based on the input SQL query



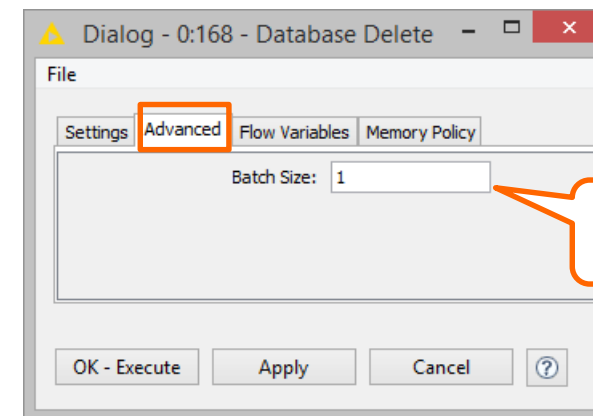
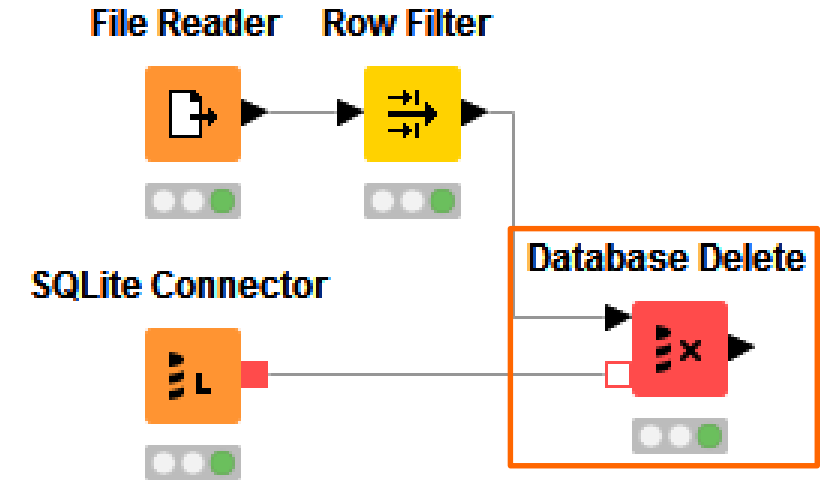
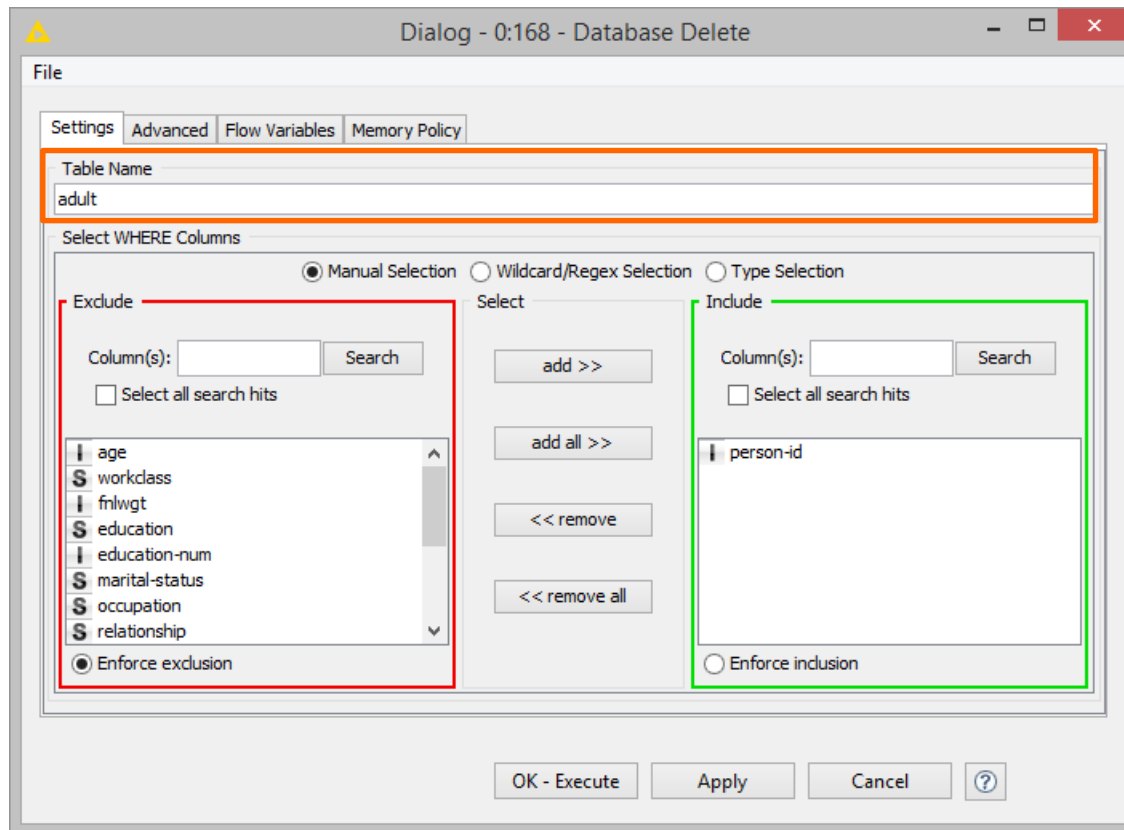
Database Update

- Updates all database records that match the update criteria



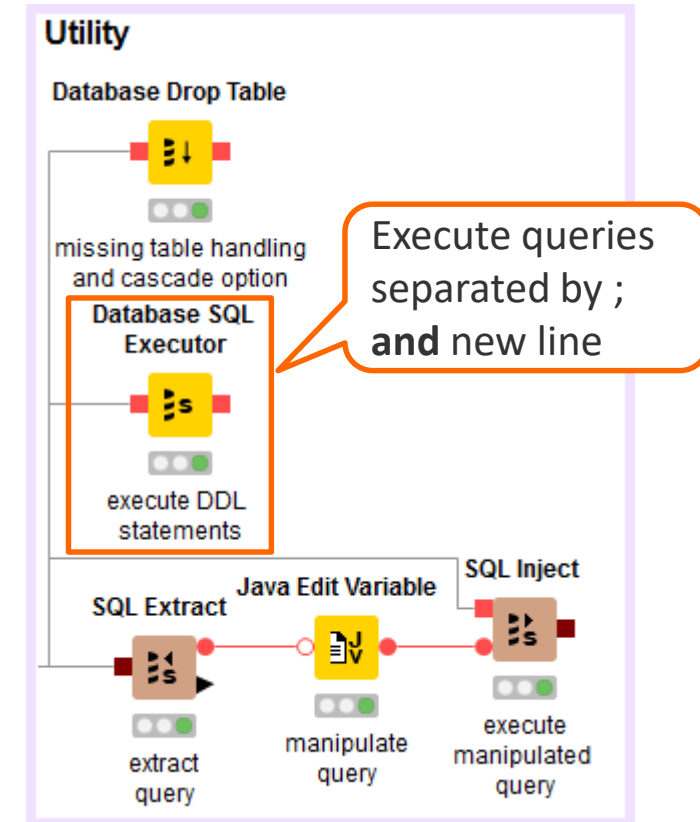
Database Delete

- Deletes all database records that match the values of the selected columns



Utility

- Drop table
 - missing table handling
 - cascade option
- Execute any SQL statement e.g. DDL
- Manipulate existing queries



Section Exercise – 04_DB_WritingToDB

- From tables ss13hme (house data) and ss13pme (person data) in database newCensus.sqlite, after joining, filtering, aggregation, prediction, timestamp creation and model conversion from PMML to table cell:
- write the original table to ss13pme_original table with a Database Connection Table Writer node ... just in case we mess up with the updates in the next step
- update all rows in ss13pme table with the output of the predictor node, that is all rows with missing COW value with the predicted COW value, using column SERIAL NO for WHERE condition (SERIAL NO uniquely identifies each person). Check the UpdateStatus column for success.

Optional: Write the learned Decision Tree Model and the timestamp into a new table named "model"

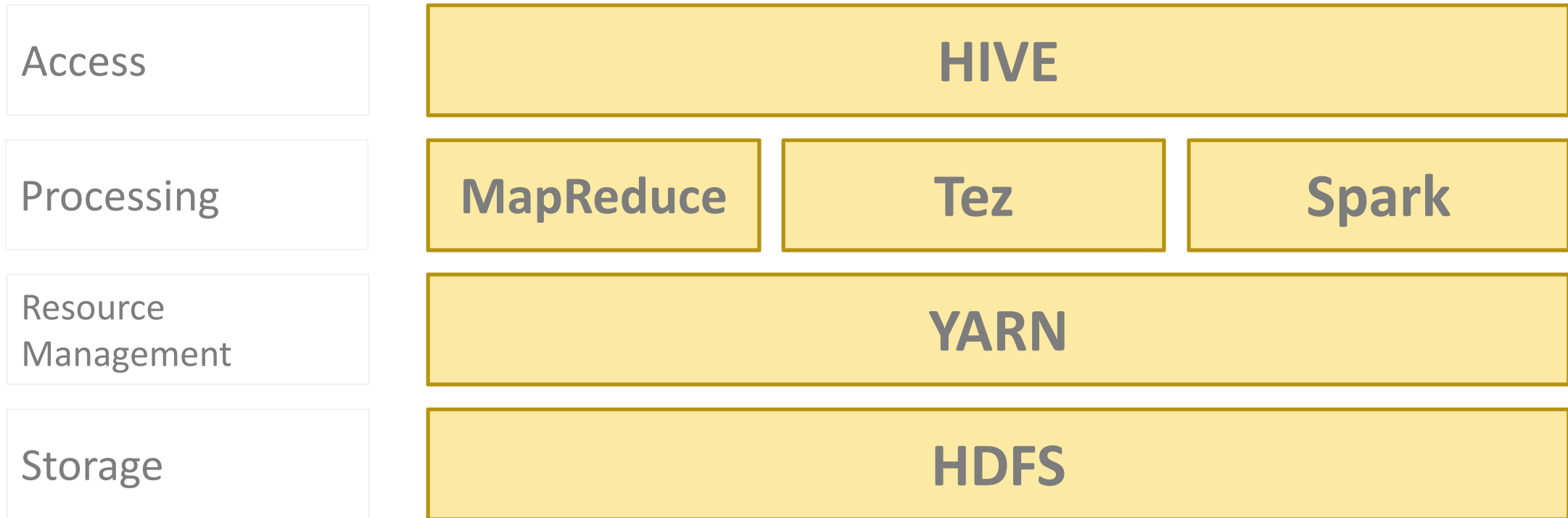
Let's try now the same with Hadoop

A quick Intro to Hadoop

Apache Hadoop

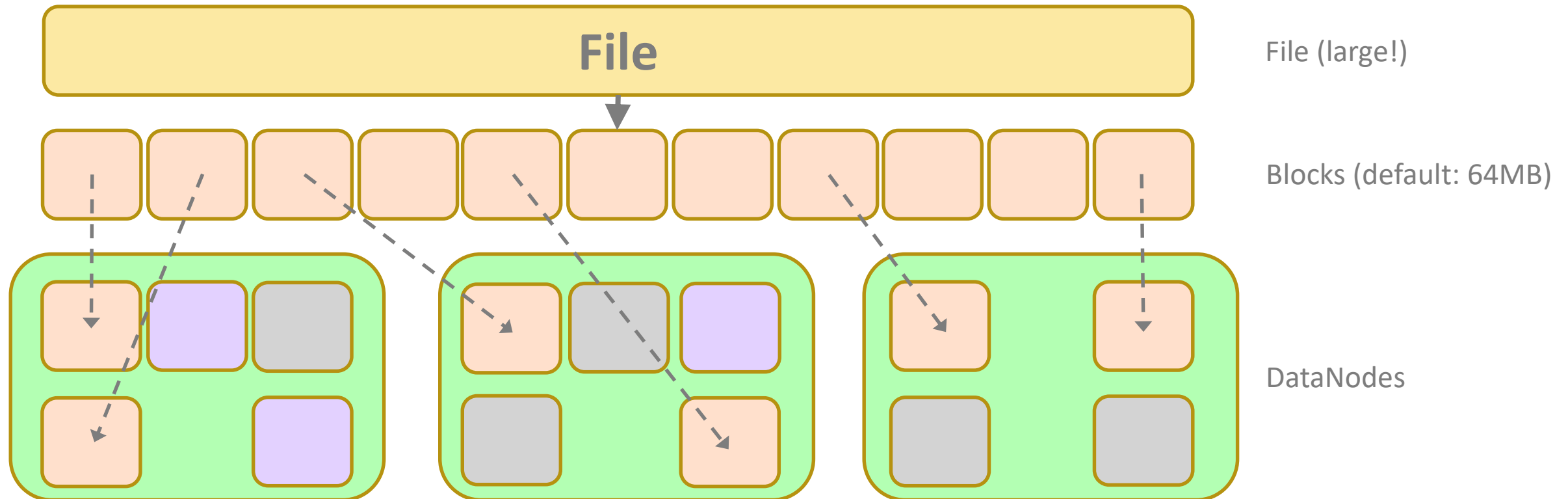
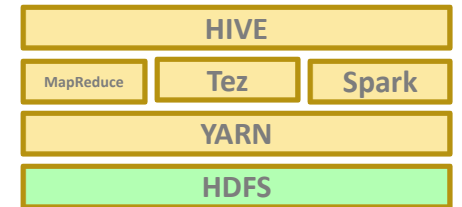
- Open-source framework for distributed storage and processing of large data sets
- Designed to scale up to thousands of machines
- Does not rely on hardware to provide high availability
 - Handles failures at application layer instead
- First release in 2006
 - Rapid adoption, promoted to top level Apache project in 2008
 - Inspired by Google File System (2003) paper
- Spawned diverse ecosystem of products

Hadoop Ecosystem



HDFS

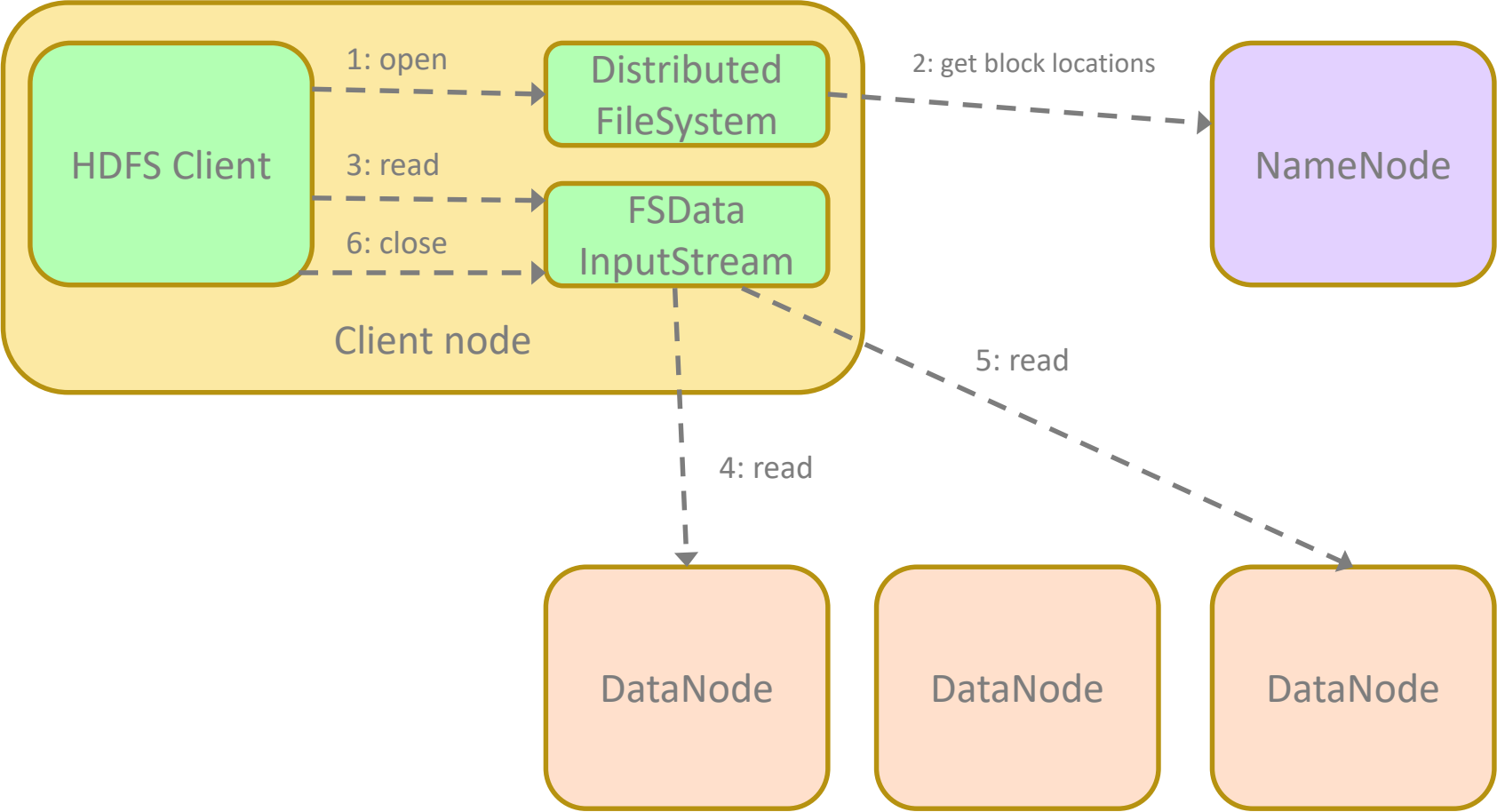
- Hadoop distributed file system
- Stores large files across multiple machines



HDFS – NameNode and DataNode

- NameNode
 - Master server that manages file system namespace
 - Maintains metadata for all files and directories in filesystem tree
 - Knows on which datanode blocks of a given file are located
 - Whole system depends on availability of NameNode
- DataNodes
 - Workers, store and retrieve blocks per request of client or namenode
 - Periodically report to namenode that they are running and which blocks they are storing

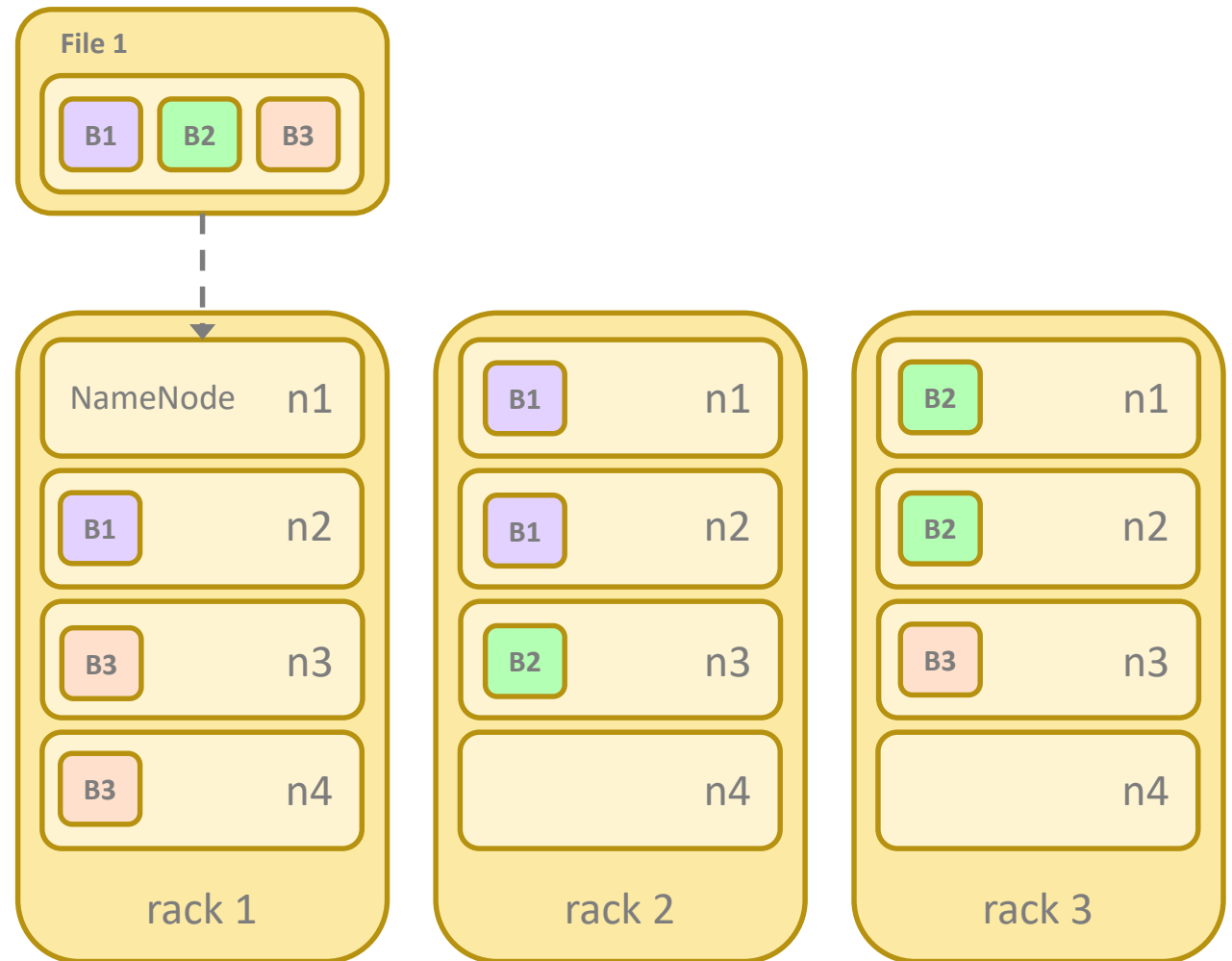
Reading Data from HDFS



HDFS – Data replication and file size

Data Replication

- All blocks of a file are stored as sequence of blocks
- Blocks of a file are replicated for fault tolerance (usually 3 replicas)
 - Aims: improve data reliability, availability, and network bandwidth utilization



HDFS – Access and File Size

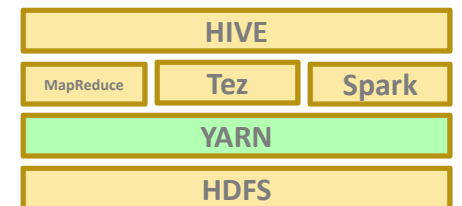
- Several ways to access HDFS data
 - FileSystem (FS) shell commands
 - Direct RPC connection
 - Requires Hadoop client to be installed
 - WebHDFS
 - Provides REST API functionality, lets external applications connect via HTTP
 - Direct transmission of data from node to client
 - Needs access to all nodes in cluster
 - HttpFS
 - All data is transmitted to client via one single node -> gateway

File Size

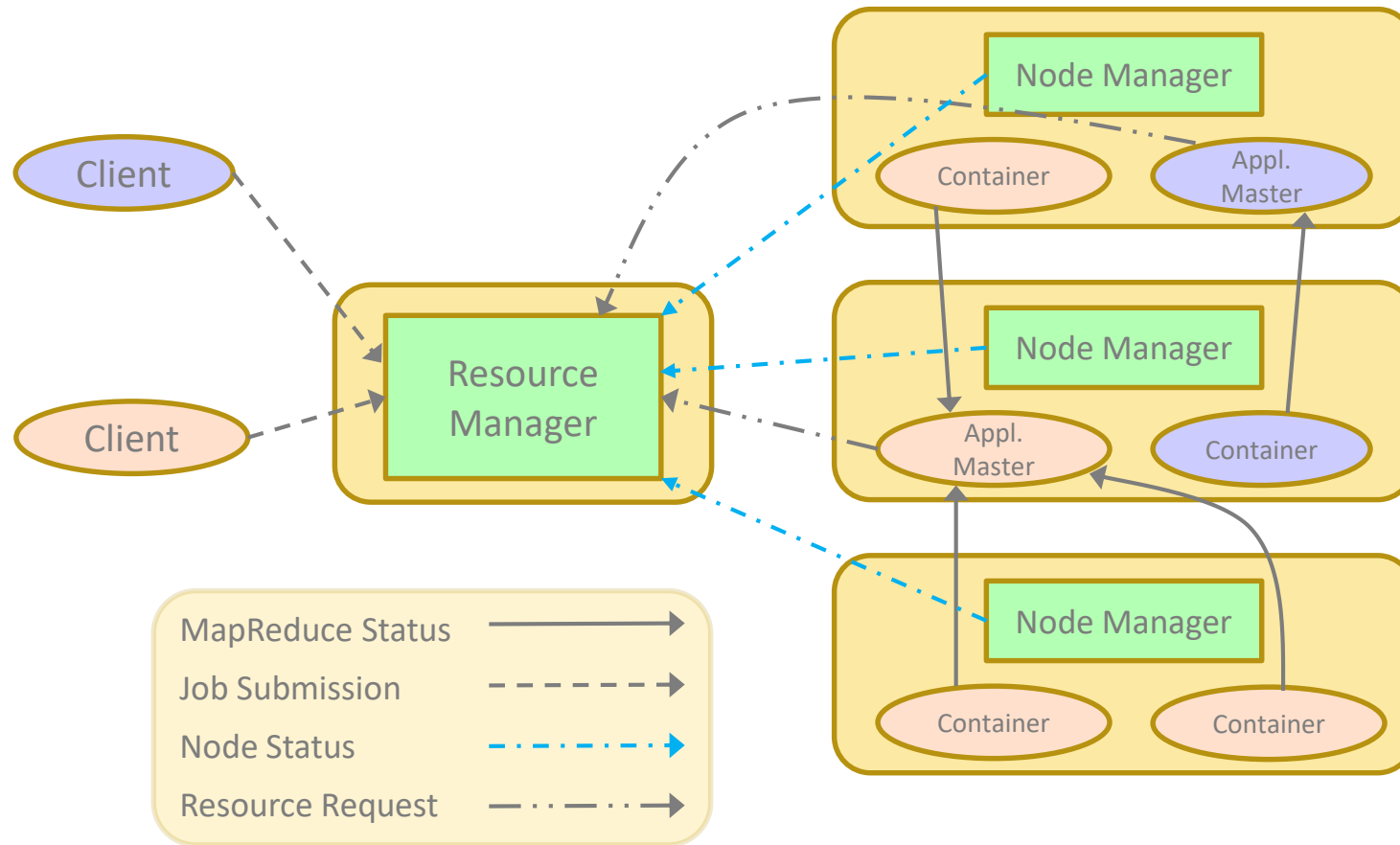
- Hadoop is designed to handle fewer large files instead of lots of small files
- Small file: File significantly smaller than Hadoop block size
- Problems:
 - Namenode memory
 - MapReduce performance

YARN

- Cluster resource management system
- Two elements
 - Resource manager (one per cluster):
 - Knows where workers are located and how many resources they have
 - Scheduler: Decides how to allocate resources to applications
 - Node manager (many per cluster):
 - Launches application containers
 - Monitor resource usage and report to Resource Manager

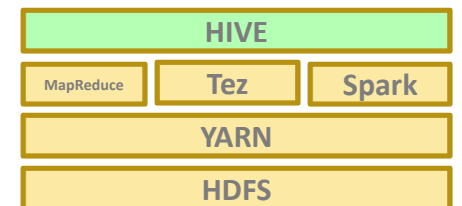


YARN



Hive

- Infrastructure on top of Hadoop
- Provides data summarization, query, and analysis
- SQL-like language (HiveQL)
- Converts queries to MapReduce, Apache Tez, and Spark jobs
- Supports various file formats:
 - Text/CSV
 - SequenceFile
 - Avro
 - ORC
 - Parquet



Spark

- Cluster computing framework for large-scale data processing
- Keeps large working datasets in memory between jobs
 - No need to always load data from disk -> way (!) faster than MapReduce
- Great for:
 - Iterative algorithms
 - Interactive analysis



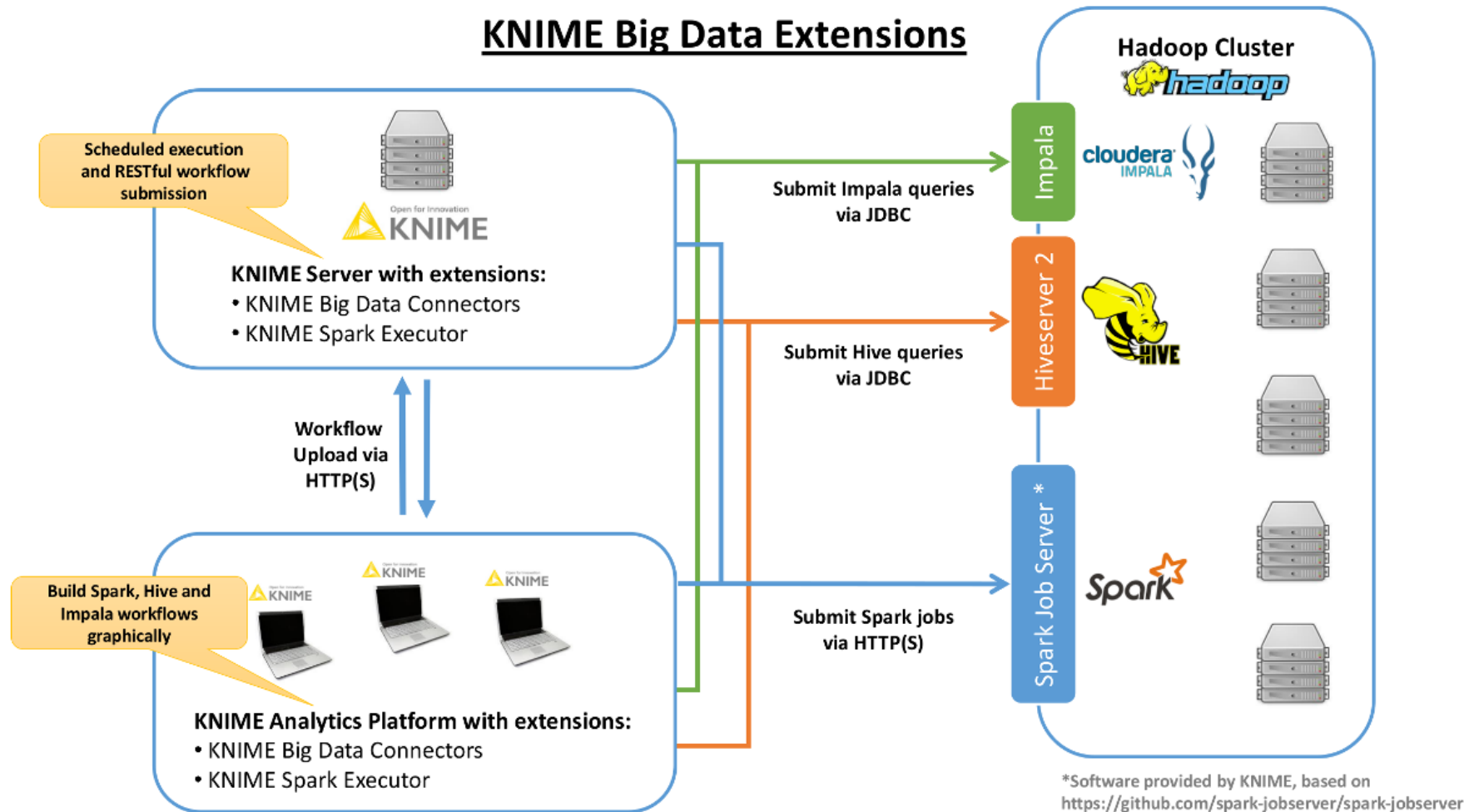
Spark – Basic Concepts

- SparkContext
 - Main entry point for Spark functionality
 - Represents connection to a Spark cluster
 - Create RDDs, accumulators, and broadcast variables on cluster
- RDD: Resilient Distributed Dataset
 - Read-only multiset of data items distributed over cluster of machines
 - Fault-tolerant: Lost partition automatically reconstructed from RDDs it was computed from
 - Lazy evaluation: Computation only happens when action is required

Spark – DataFrame and Dataset

- DataFrame
 - Distributed collection of data organized in named columns
 - Similar to table in relational database
 - Can be constructed from many sources: structured data files, Hive table, RDDs...
- Dataset
 - Extension of DataFrame API
 - Strongly-typed, immutable collection of objects mapped to a relational schema
 - Catches syntax and analysis errors at compile time

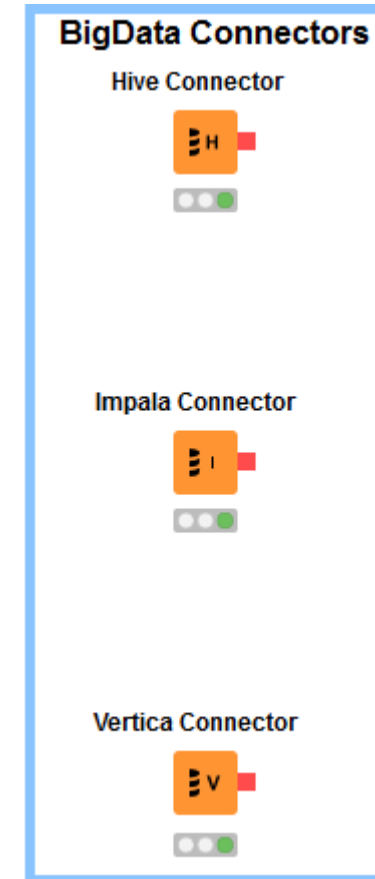
Hive, HDFS, Spark Architecture



In-Database Processing on Hadoop

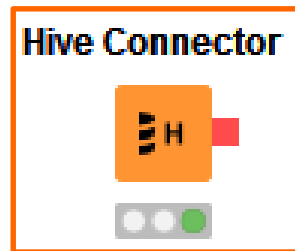
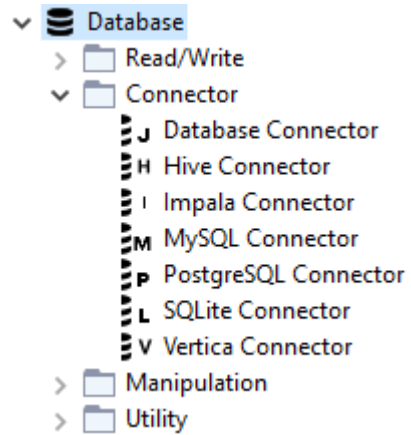
KNIME Big Data Connectors

- Package required drivers/libraries for specific HDFS, Hive, Impala access
- Preconfigured connectors
 - Hive
 - Impala



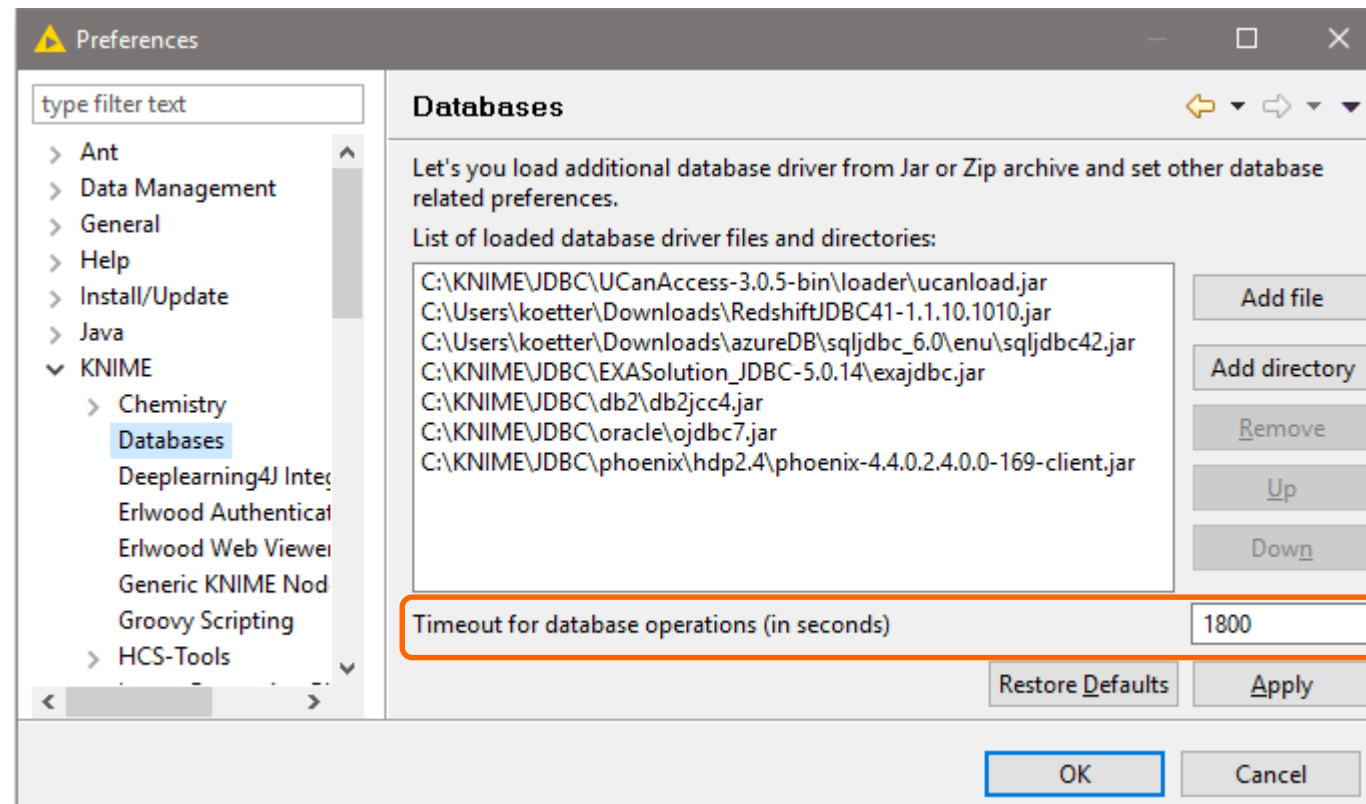
Hive Connector

- Creates JDBC connect string to connect to Hive db
- On unsecured clusters no password required

A screenshot of the 'Dialog - 0:131 - Hive Connector' configuration window. The window has a title bar with a yellow triangle icon and standard window controls. Below the title bar is a 'File' menu. The main content area is divided into three tabs: 'Flow Variables', 'Job Manager Selection', and 'Memory Policy'. The 'Job Manager Selection' tab is active, showing 'Connection settings'. Under 'Connection', there are fields for 'Hostname' (sandbox.hortonworks.com), 'Port' (10.000), 'Database name', and 'Parameter'. Under 'Authentication', the 'Use username & password' option is selected, with fields for 'Username' (knime) and 'Password'. Under 'Timezone correction', the 'Use local timezone' option is selected, with a dropdown for 'Europe/Berlin'. Under 'Misc', there are checkboxes for 'Validate connection on close' and 'Retrieve metadata in configure'. At the bottom are buttons for 'OK', 'Apply', 'Cancel', and a help icon.

Preferences

- Time till timeout has to be longer than usual when using Hadoop Hive (data retrieval time might be long)



Section Exercise – 0123_Hive_Modelling

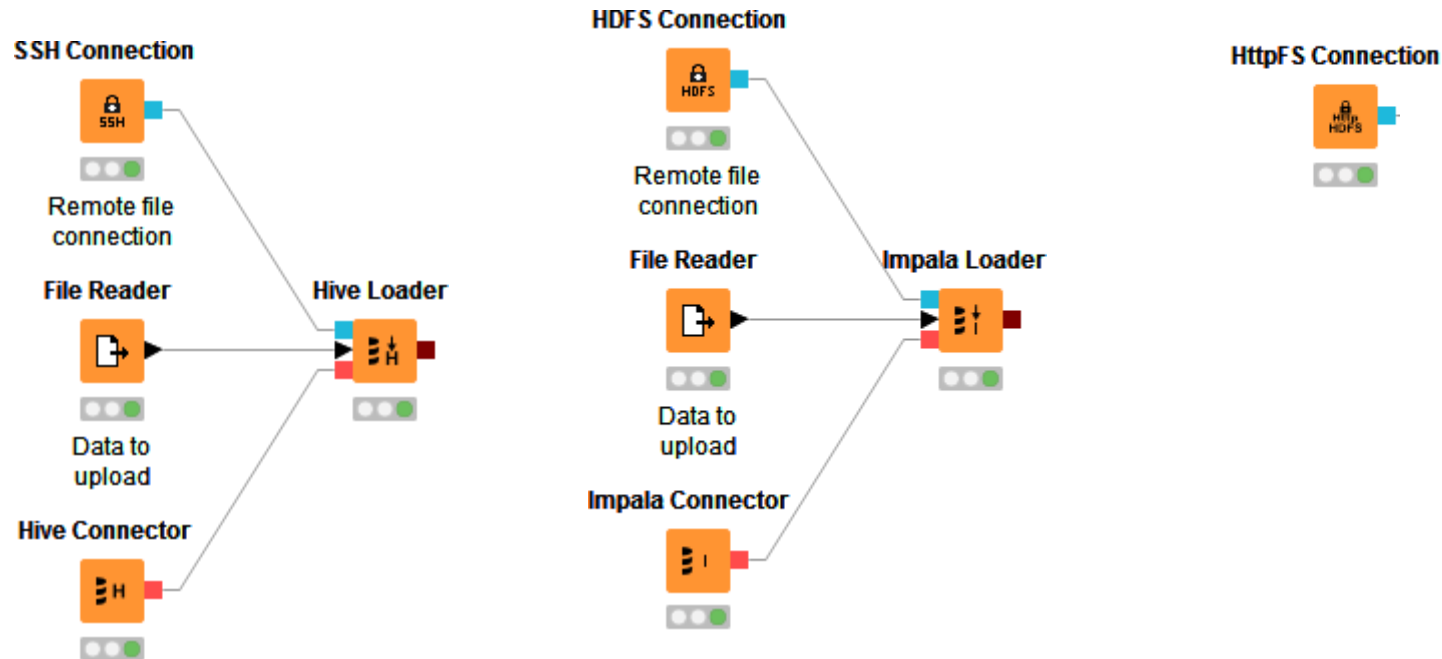
On the workflow implemented in the previous section to predict missing COW values, move execution from database to Hive. That is:

- change this workflow to run on the ss13pme table on the Hive database
 - Hive URL: *see handout*
 - Username: *see handout*
 - no password
- **Warning.** Concurrent access to Hive might generate an error. Use flow variable connections to generate execution dependencies among nodes

Write/Load Data into Hadoop

Hive Loader

- Upload a KNIME data table to Hive/Impala
- Part of the commercial KNIME Big Data Connectors Extension



HttpFS Connection

- Connect to HDFS
- Needs user and machine URL and port
- Output port is squared blue like SSH Connection node

HttpFS Connection



Dialog - 4:195 - HttpFS Connection (connec... — □ ×

File

Options Flow Variables Job Manager Selection Memory Policy

Host: ec2-54-73-143-134.eu-west-1.compute.amazonaws.com

Port: 14,000

Authentication
☒ User ☐ Kerberos

Workflow credentials
☐ [Dropdown]

User: hadoop

Timeout: 30,000

Test connection

OK Apply Cancel ?

Hive Loader

Dialog - 4:192 - Hive Loader (create/override)

File

Load settings | SQL Types | Flow Variables | Job Manager Selection | Memory Policy

Target folder: /tmp/ Browse...

Table name: newTable

☒ Drop existing table

Partition columns

Exclude

Column(s): Search

☐ Select all search hits

rt
sporder
st
pwgtp
agep
cit
citwp05
citwp12

Select

add >>

add all >>

<< remove

<< remove all

Include

Column(s): Search

☐ Select all search hits

schl

OK Apply Cancel ?

Partitioning influences performance. Partition columns shouldn't contain missing values

Section Exercise – 04_Hive_WritingToDB

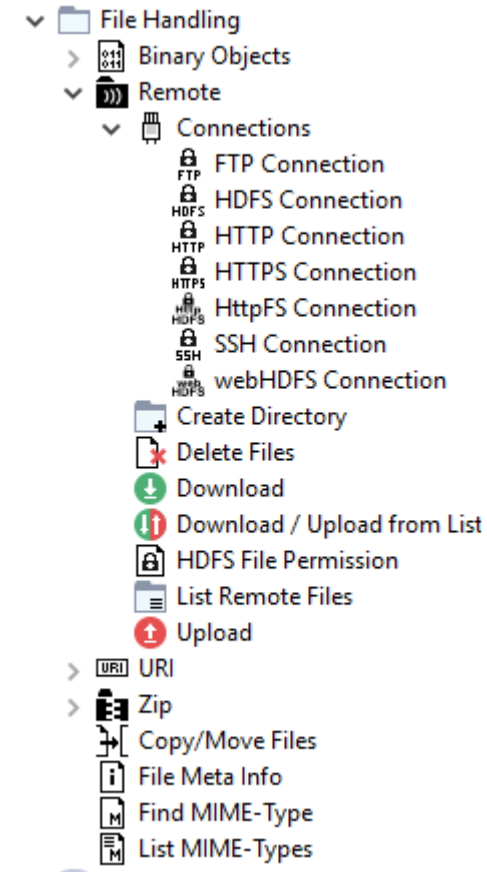
Start from the workflow that implements the missing value strategy and write the results back into Hive. That is:

- Write the results onto a new table in Hive using an HttpFS Connection node and a Hive Loader node
 - New table name: *see handout*
 - Hive URL: *see handout*
 - Username: *see handout*
 - no password

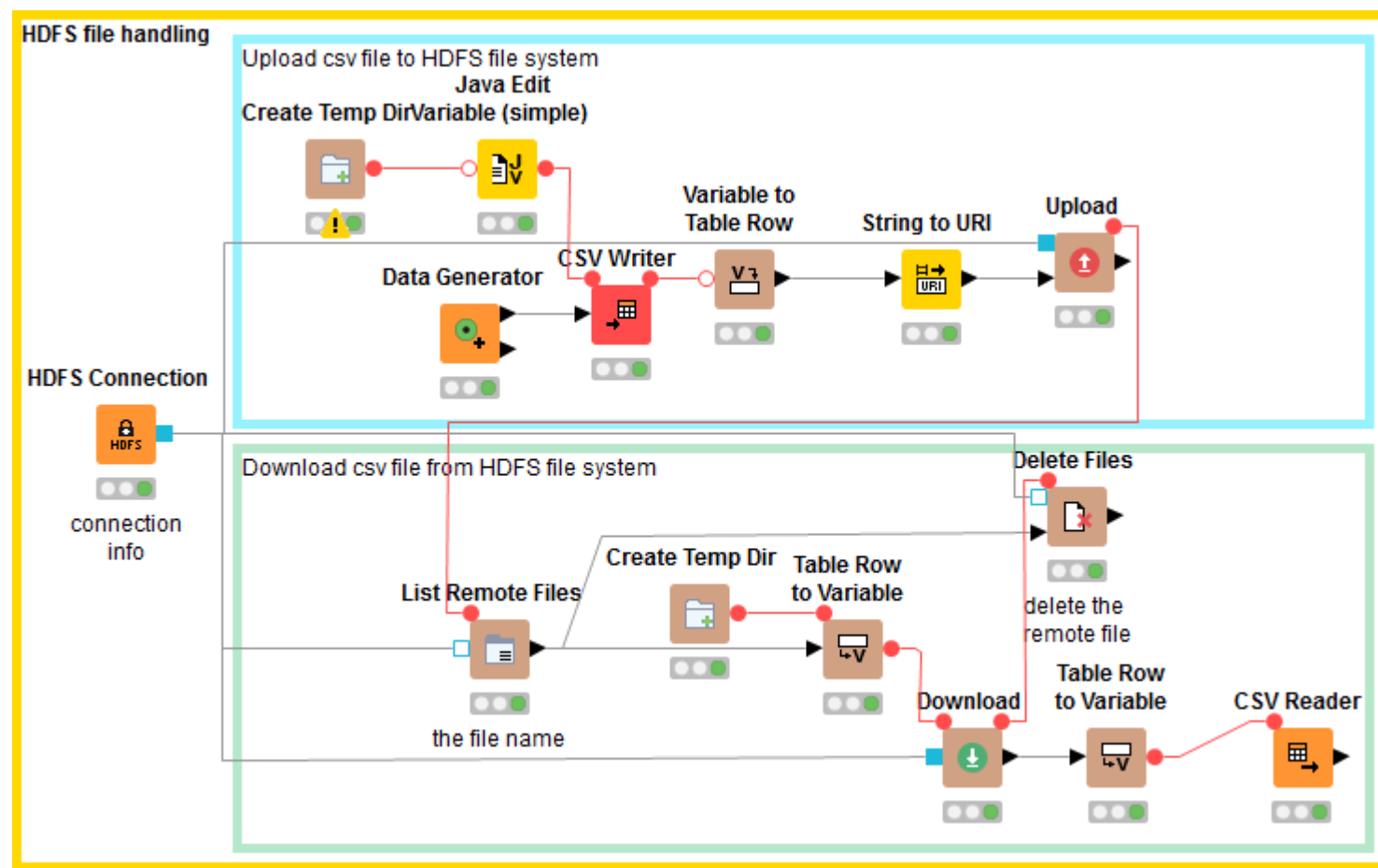
HDFS File Handling

HDFS File Handling

- New nodes
 - HDFS/HttpFS/webHDFS Connection
 - HDFS File Permission
- Utilize the existing remote file handling nodes
 - Upload/download files
 - Create/list directories
 - Delete files

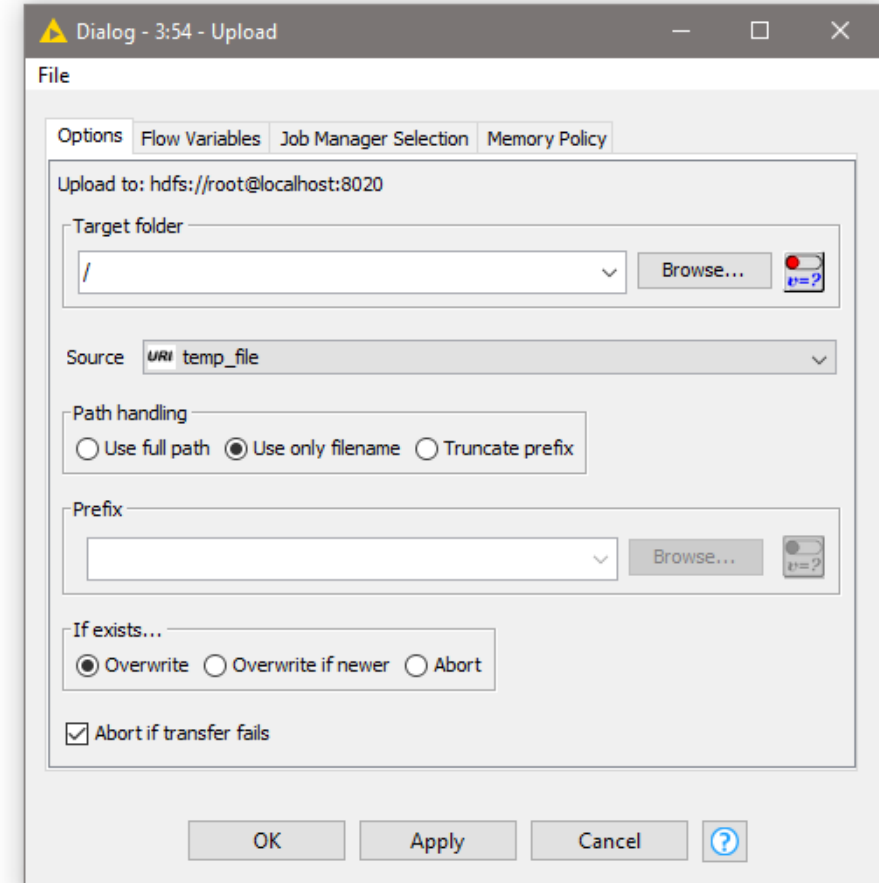
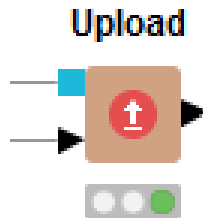


HDFS File Handling



Upload

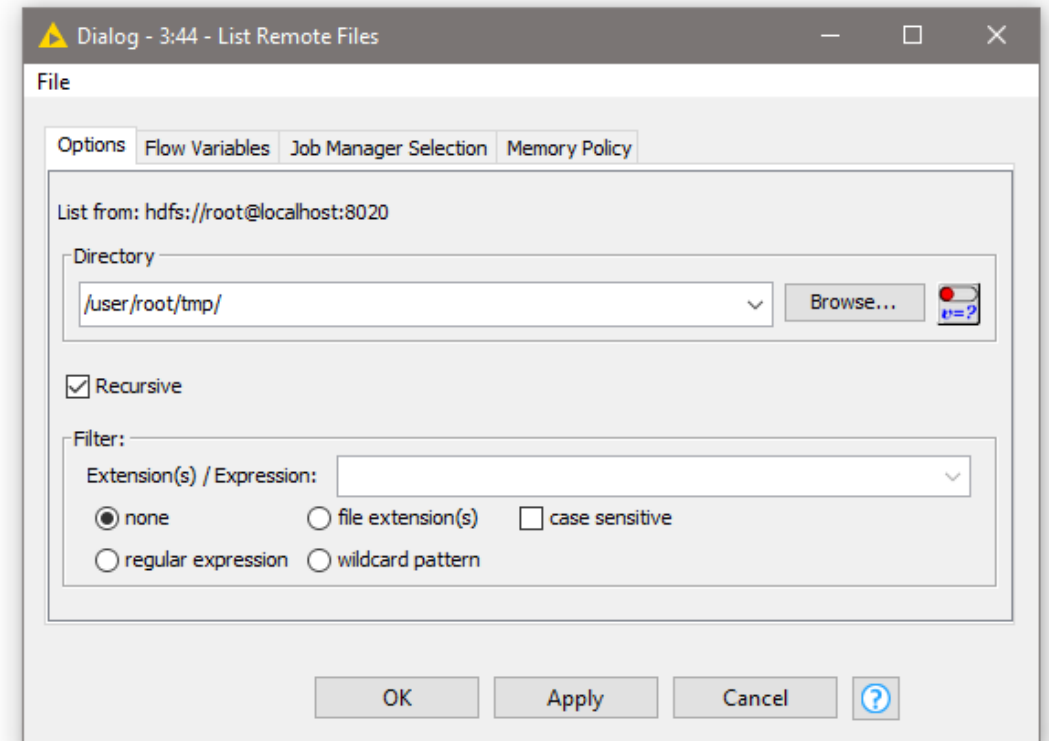
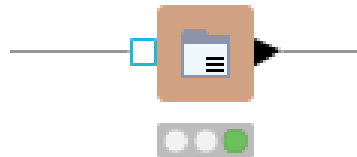
- From the connection to HDFS, uploads a file from a local URL to a target folder on HDFS



List Remote Files

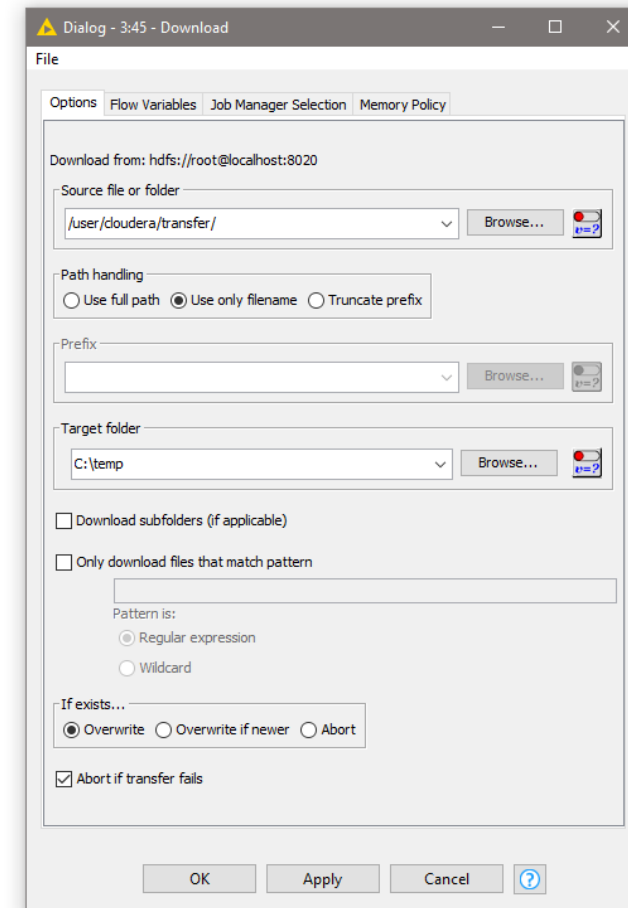
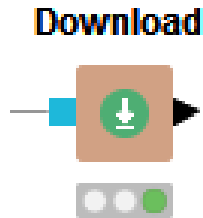
- Lists all Files in folder on HDFS
- Recursive option and file extension filtering

List Remote Files



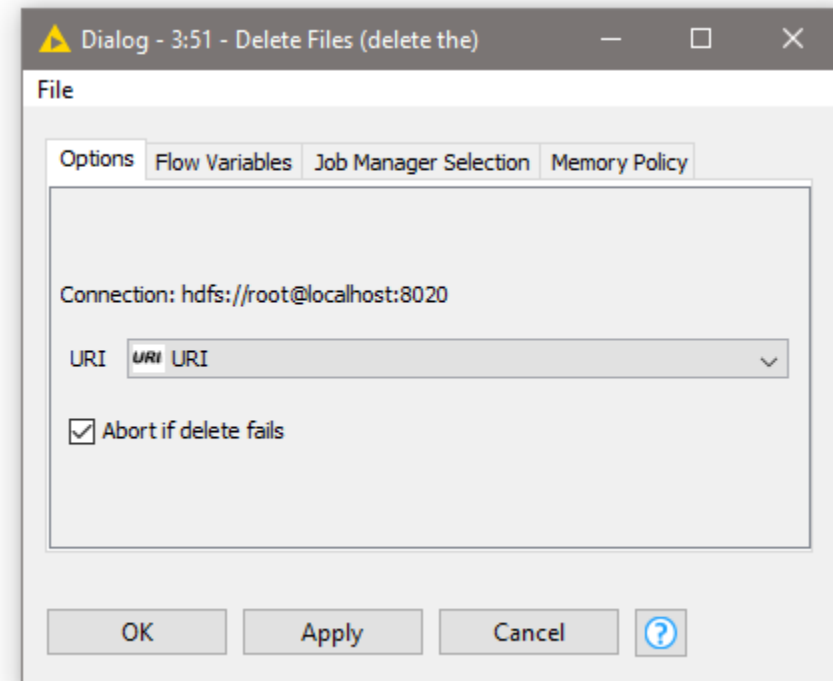
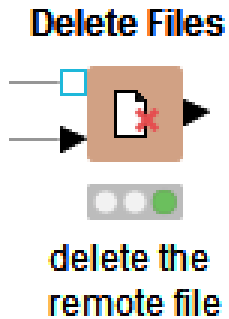
Download

- Downloads file from HDFS to a local directory recursively (if chosen)



Delete Files

- Delete Files from URI on HDFS connection



Pre-processing on Hadoop

- Case Study –

Pre-processing for Energy Usage Prediction

Energy Usage Prediction from Smart Meters Data

- Read Smart Meter Energy Data
- Clean Up and Aggregate total Energy Usage by hour, week, day, month, year
- Calculate Behavioral Measures for each Smart Meter

Workflow 1

- Cluster Smart Meters with Similar Behavior (k-Means)

Workflow 2

Not part of this training

- Predict Energy Usage in Clustered Smart Meters (Auto-Regressive Time Series Prediction)

Workflow 3

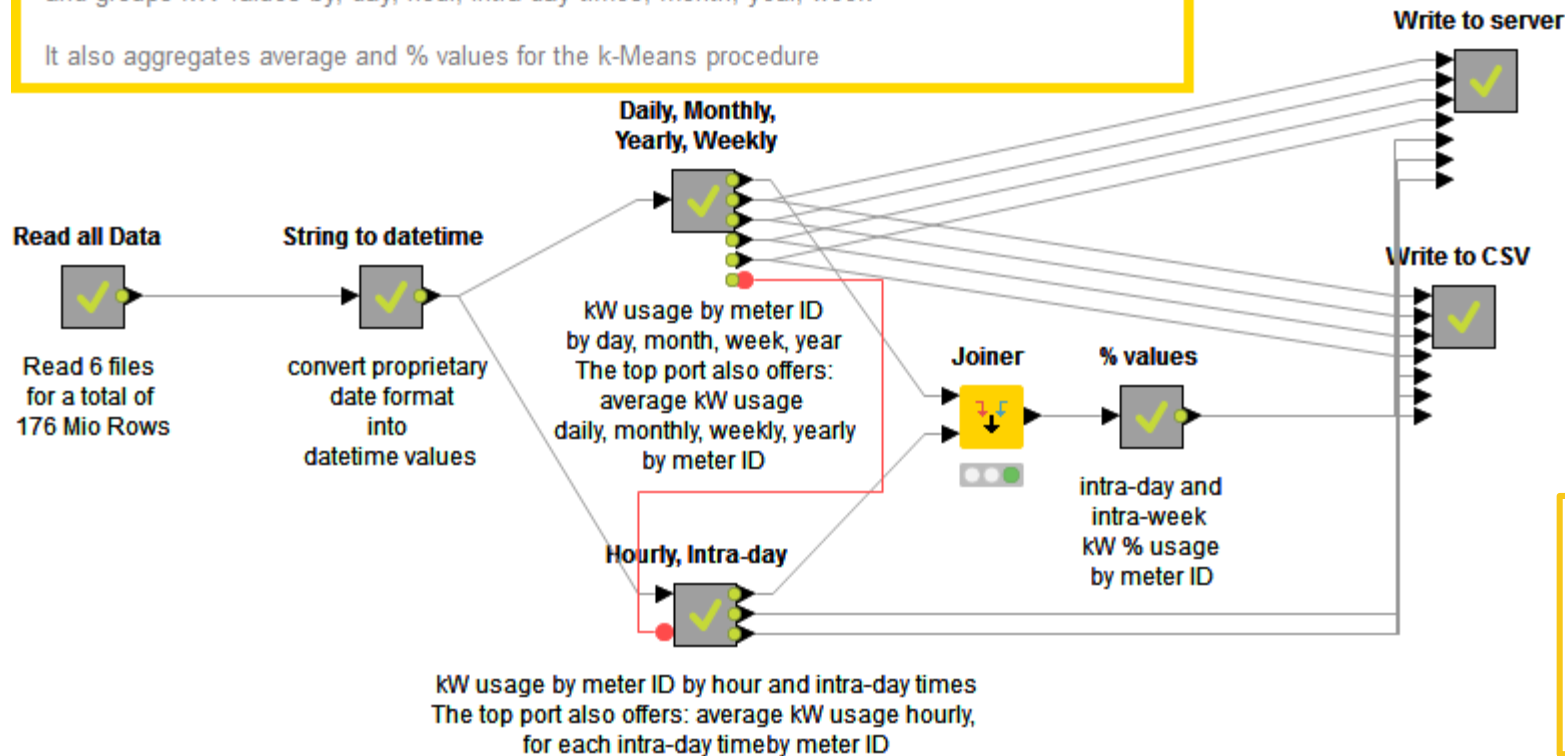
Not part of this training

Workflow 1: PrepareData (in KNIME)

Energy Usage Prediction from Smart Meters Data

This workflow reads Ireland's electricity data, converts the dates from the proprietary format into datetime values, and groups kW values by day, hour, intra-day times, month, year, week

It also aggregates average and % values for the k-Means procedure



**Runtime:
~2 days**

Irish Smart Energy Meter Trials

- July 2009 – Dec 2010
- 6000 meters
- roughly 176m rows of data

Workflow 1: PrepareData (In-Database Processing)

Energy Usage Prediction from Smart Meters Data

This workflow reads Ireland's electricity data, converts the dates from the proprietary format into datetime values, and groups kW values by day, hour, intra-day times, month, year, week

It also aggregates average and % values for the k-Means procedure

Connect to a big data platform:

- Impala
 - Hive
 - parStream
- to read the energy data

Database Connector



parStream platform

Impala Connector



Cloudera Amazon cluster

Hive Connector



Hive Cluster

Database Table Selector



table "energy" with all energy measurements sampled every half an hour one year long almost 6000 meter IDs

String to datetime



change datetime format

SQL queries to calculate Energy Usage Measures

daily, hourly, weekly, monthly, yearly, by intra-day segments

Daily, Monthly, Yearly, Weekly



kW usage by meter ID by day, month, week, year
The top port also offers:
average kW usage
daily, monthly, weekly, yearly by meter ID

Hourly, Intra-day



kW usage by meter ID by hour and intra-day times
The top port also offers:
average kW usage hourly, for each intra-day time by meter ID

Database Joiner



join time series

% values



intra-day and intra-week kW % usage by meter ID

Database Connection Table Reader



monthly
Database Connection Table Reader



daily
Database Connection Table Reader



yearly
Database Connection Table Reader



weekly
Database Connection Table Reader



result
Database Connection Table Reader



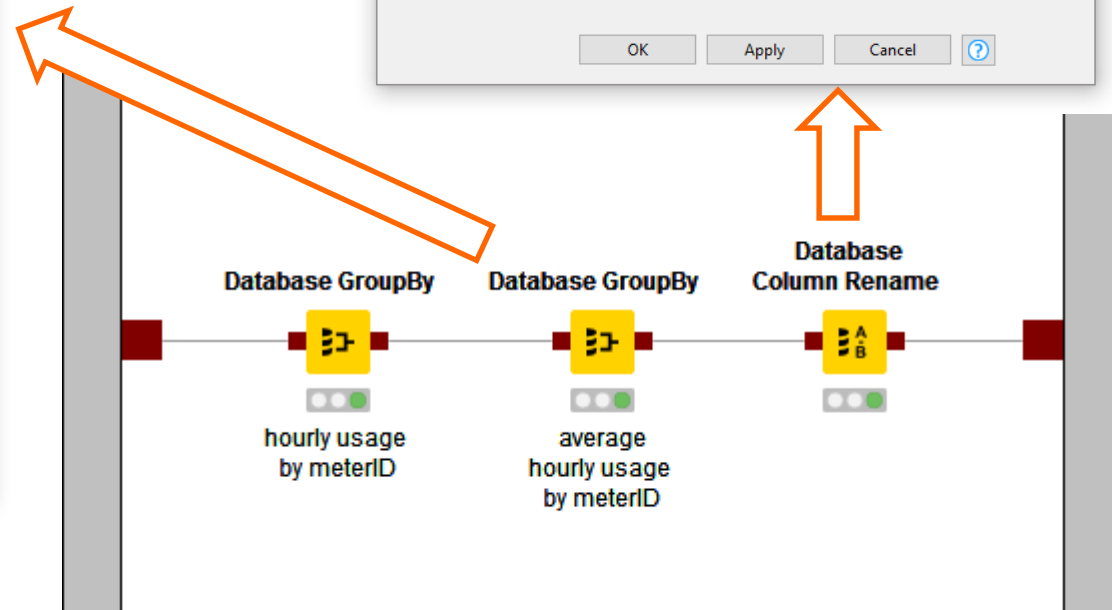
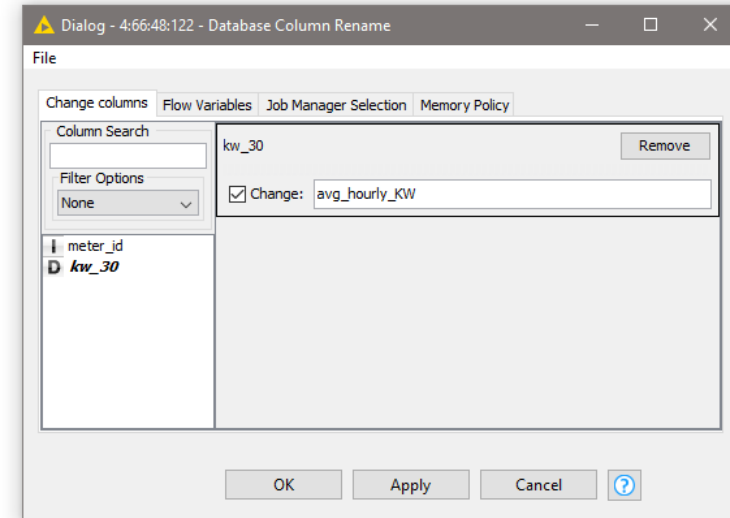
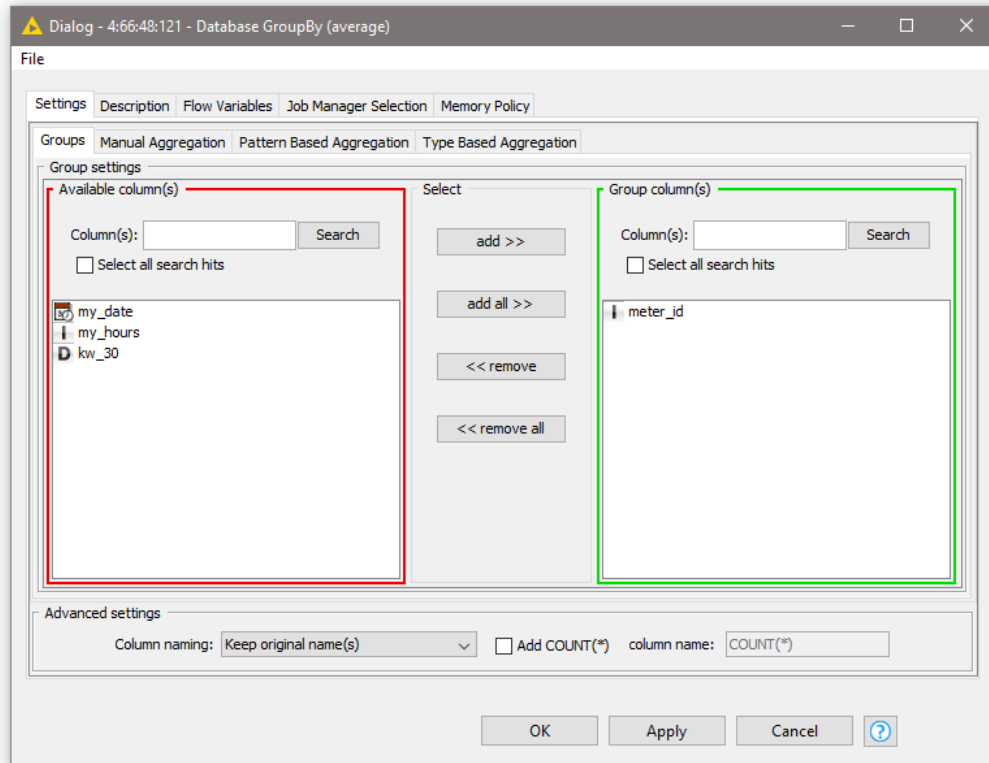
hourly
Database Connection Table Reader

Run the SQL queries and retrieve final data

Irish Smart Energy Meter Trials

- July 2009 – Dec 2010
- 6000 meters
- roughly 176m rows of data

Average Hourly Values



Import Aggregated Data from Database into KNIME

Energy Usage Prediction from Smart Meters Data

This workflow reads Ireland's electricity data, converts the dates from the proprietary format into datetime values, and groups kW values by day, hour, intra-day times, month, year, week

It also aggregates average and % values for the k-Means procedure

Connect to a big data platform:

- Impala
 - Hive
 - parStream
- to read the energy data

Database Connector



parStream platform

Impala Connector



Cloudera Amazon cluster

Hive Connector



Hive Cluster

Database Table Selector



table "energy" with all energy measurements sampled every half an hour one year long almost 6000 meter IDs

String to datetime



change datetime format

SQL queries to calculate Energy Usage Measures

daily, hourly, weekly, monthly, yearly, by intra-day segments

Daily, Monthly, Yearly, Weekly



kW usage by meter ID by day, month, week, year
The top port also offers:
average kW usage daily, monthly, weekly, yearly by meter ID

Hourly, Intra-day



kW usage by meter ID by hour and intra-day times
The top port also offers:
average kW usage hourly, for each intra-day time by meter ID

Database Joiner



join time series

% values



intra-day and intra-week kW % usage by meter ID

Database Connection Table Reader



monthly
Database Connection Table Reader



daily
Database Connection Table Reader



yearly
Database Connection Table Reader



weekly
Database Connection Table Reader



result
Database Connection Table Reader



hourly
Database Connection Table Reader

Run the SQL queries and retrieve final data

**Runtime:
< 30 min**

Irish Smart Energy Meter Trials

- July 2009 – Dec 2010
- 6000 meters
- roughly 176m rows of data

Ready for Spark?

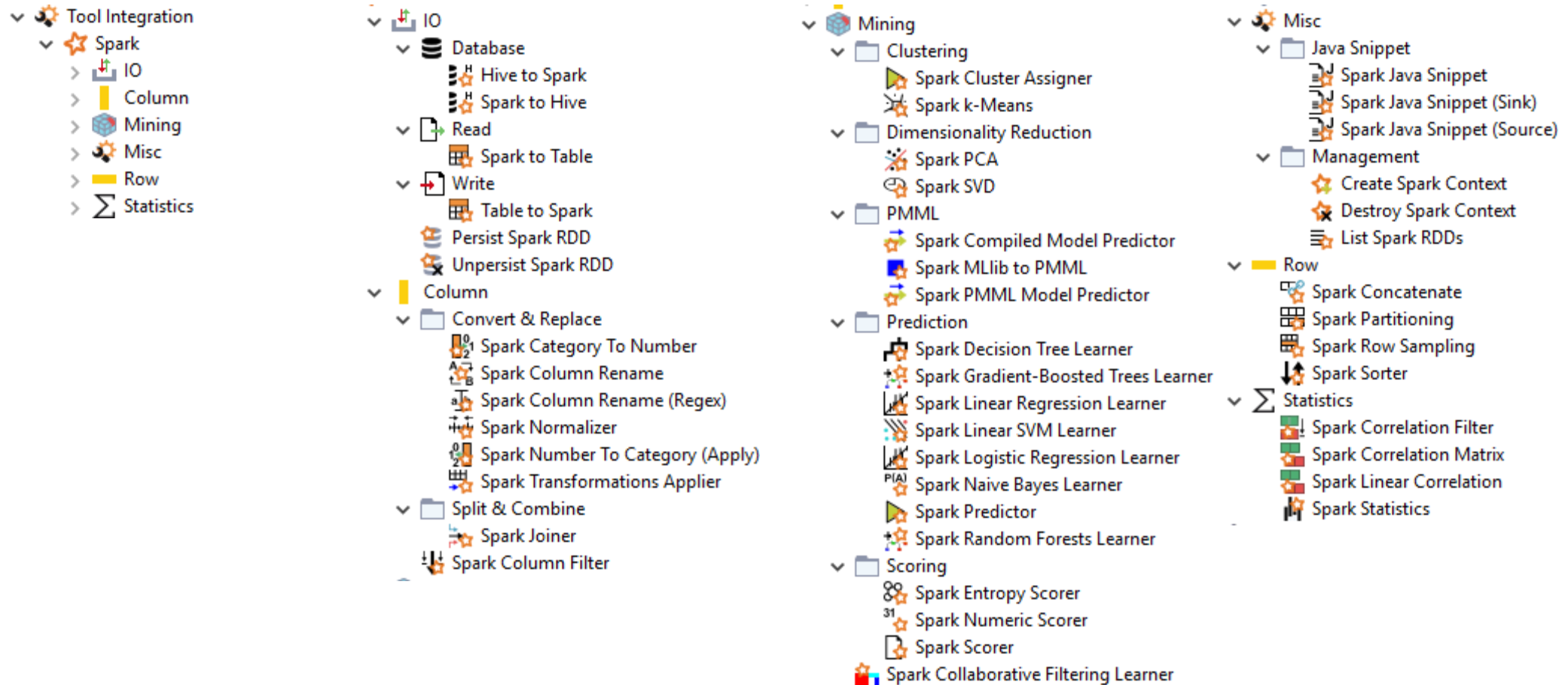
KNIME Spark Executor

Spark: Machine Learning on Hadoop

- Runs on Hadoop
- Supported Spark Versions
 - 1.2, 1.3, 1.5, 1.6
- One KNIME Spark Executor for all Spark versions
- Scalable machine learning library (Spark MLlib)
- Algorithms for
 - Classification (decision tree, naïve bayes, ...)
 - Regression (logistic regression, linear regression, ...)
 - Clustering (k-means)
 - Collaborative filtering (ALS)
 - Dimensionality reduction (SVD, PCA)



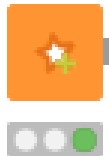
Spark Integration in KNIME



Create/Destroy Spark Context

- Create a new Spark context
- Changes KNIME settings for a workflow branch
- Destroying Spark Context destroys all Spark RDDs within the context

Create Spark Context



Destroy Spark Context



Dialog - 0:132 - Create Spark Context

File

Flow Variables | Job Manager Selection | Memory Policy

Context Settings | Connection Settings

Spark version: 1.6

Context name: knimeSparkContext

☒ Delete objects on dispose

Spark job log level:
☐ DEBUG ☐ INFO ☒ WARN ☐ ERROR

☐ Override spark settings

Custom spark settings:
memory-per-node: 512m

OK Apply Cancel ?

Dialog - 0:132 - Create Spark Context

File

Flow Variables | Job Manager Selection | Memory Policy

Context Settings | Connection Settings

Job server URL: http://localhost:8090/

Authentication
☒ None
☐ Credentials
☐ Username & password

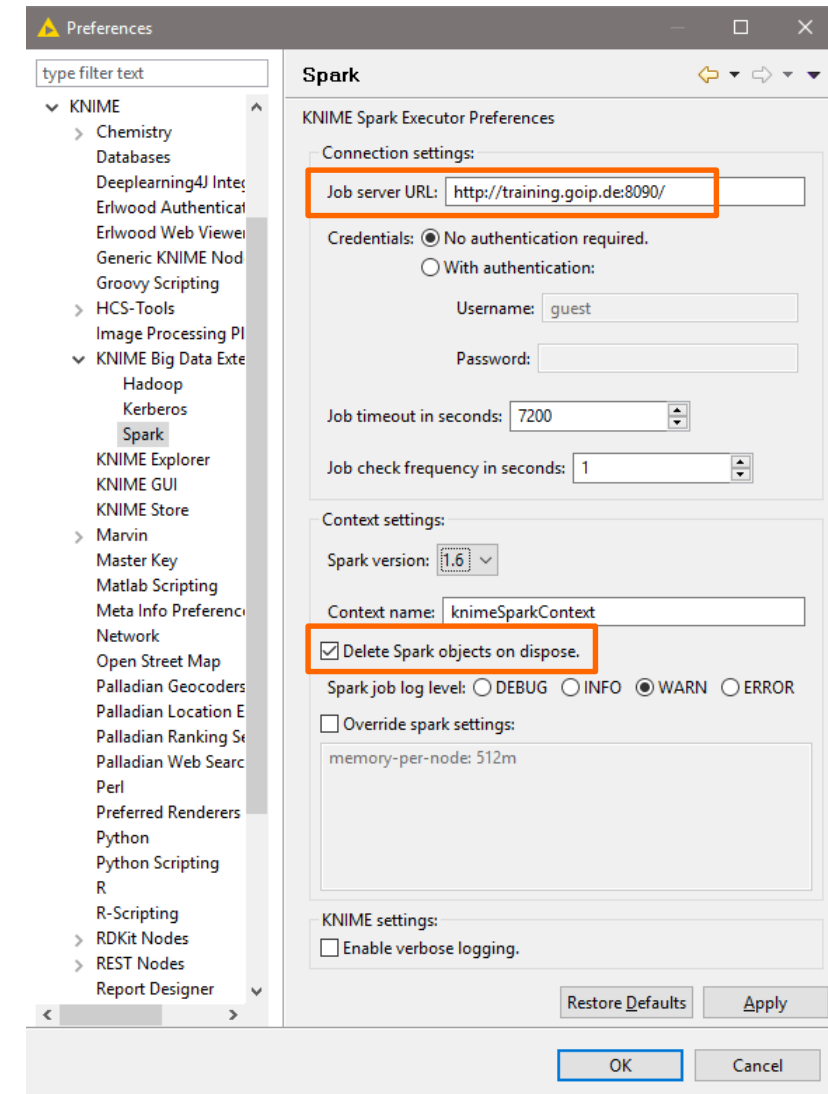
Spark job timeout (seconds): 7.200

Spark job check frequency (seconds): 1

OK Apply Cancel ?

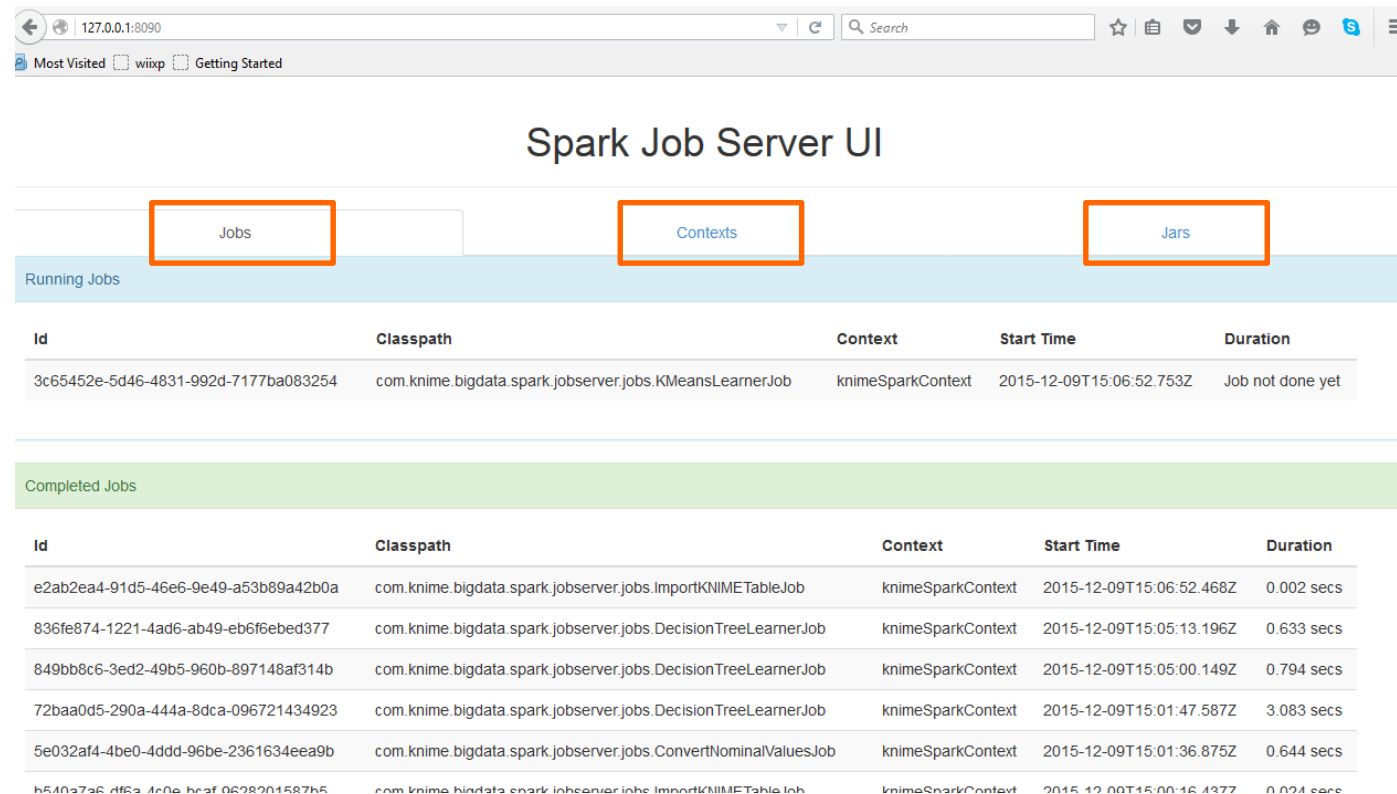
KNIME Spark Preferences: Default Spark Context

- Connection settings
 - Job server URL
 - Authentication
- Set job time outs
- Context Settings
 - Spark version
 - Spark RDD handling
 - Log level
 - Additional Spark settings



Spark Job Server Console

- `http://<server-url>:8090/` to connect to Spark job server UI



The screenshot shows the Spark Job Server UI in a web browser. The address bar shows the URL `http://127.0.0.1:8090/`. The page title is "Spark Job Server UI". There are three tabs: "Jobs", "Contexts", and "Jars". The "Jobs" tab is selected and highlighted with an orange box. Below the tabs, there are two sections: "Running Jobs" and "Completed Jobs".

Running Jobs

Id	Classpath	Context	Start Time	Duration
3c65452e-5d46-4831-992d-7177ba083254	com.knime.bigdata.spark.jobserver.jobs.KMeansLearnerJob	knimeSparkContext	2015-12-09T15:06:52.753Z	Job not done yet

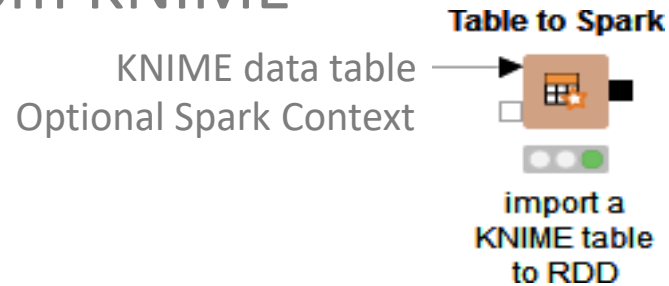
Completed Jobs

Id	Classpath	Context	Start Time	Duration
e2ab2ea4-91d5-46e6-9e49-a53b89a42b0a	com.knime.bigdata.spark.jobserver.jobs.ImportKNIMETableJob	knimeSparkContext	2015-12-09T15:06:52.468Z	0.002 secs
836fe874-1221-4ad6-ab49-eb6f6ebd377	com.knime.bigdata.spark.jobserver.jobs.DecisionTreeLearnerJob	knimeSparkContext	2015-12-09T15:05:13.196Z	0.633 secs
849bb8c6-3ed2-49b5-960b-897148af314b	com.knime.bigdata.spark.jobserver.jobs.DecisionTreeLearnerJob	knimeSparkContext	2015-12-09T15:05:00.149Z	0.794 secs
72baa0d5-290a-444a-8dca-096721434923	com.knime.bigdata.spark.jobserver.jobs.DecisionTreeLearnerJob	knimeSparkContext	2015-12-09T15:01:47.587Z	3.083 secs
5e032af4-4be0-4ddd-96be-2361634eea9b	com.knime.bigdata.spark.jobserver.jobs.ConvertNominalValuesJob	knimeSparkContext	2015-12-09T15:01:36.875Z	0.644 secs
b540a7a6-df6a-4c0a-bcaf-0628201587b5	com.knime.bigdata.spark.jobserver.jobs.ImportKNIMETableJob	knimeSparkContext	2015-12-09T15:00:16.437Z	0.024 secs

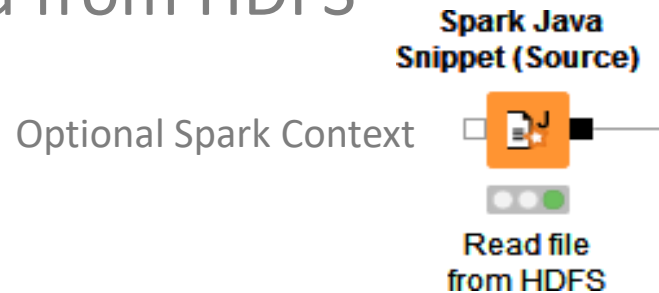
Import Data from KNIME or Hadoop

Import Data from KNIME/Hadoop to Spark

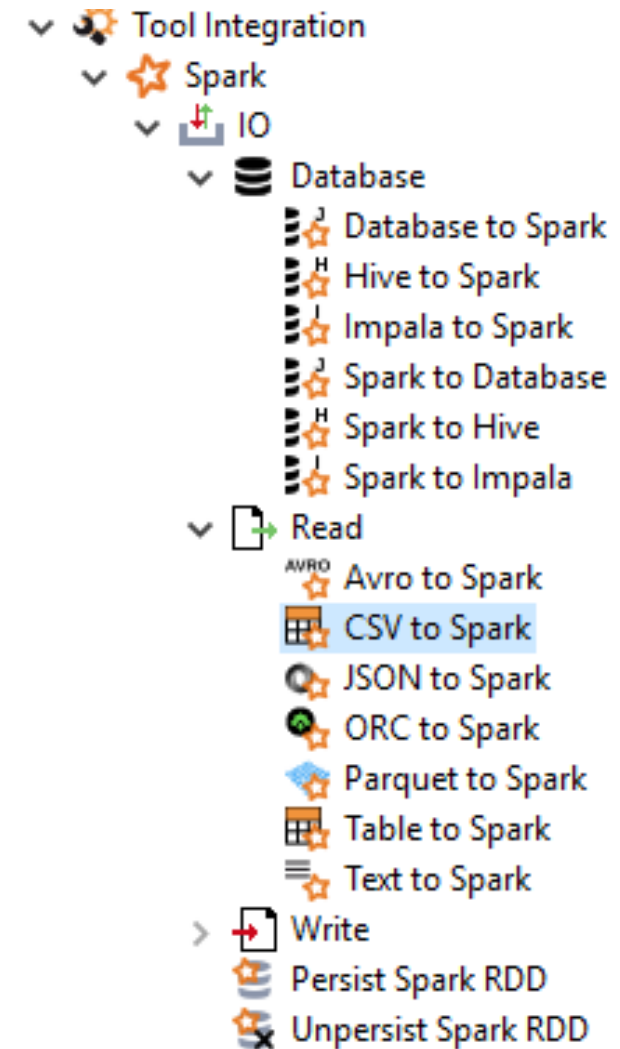
- From KNIME



- Read from HDFS



- From Hive

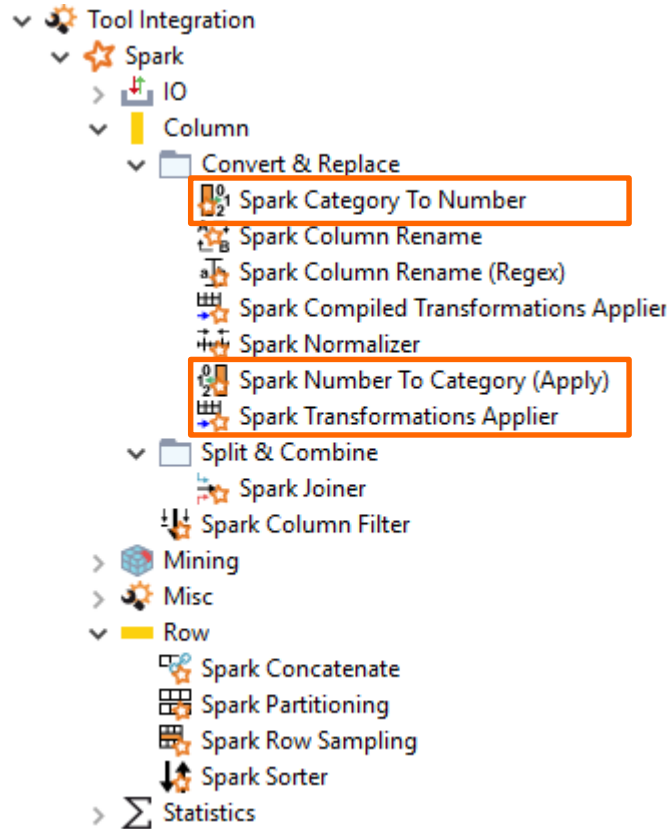


Section Exercise – 01_Spark_Connect

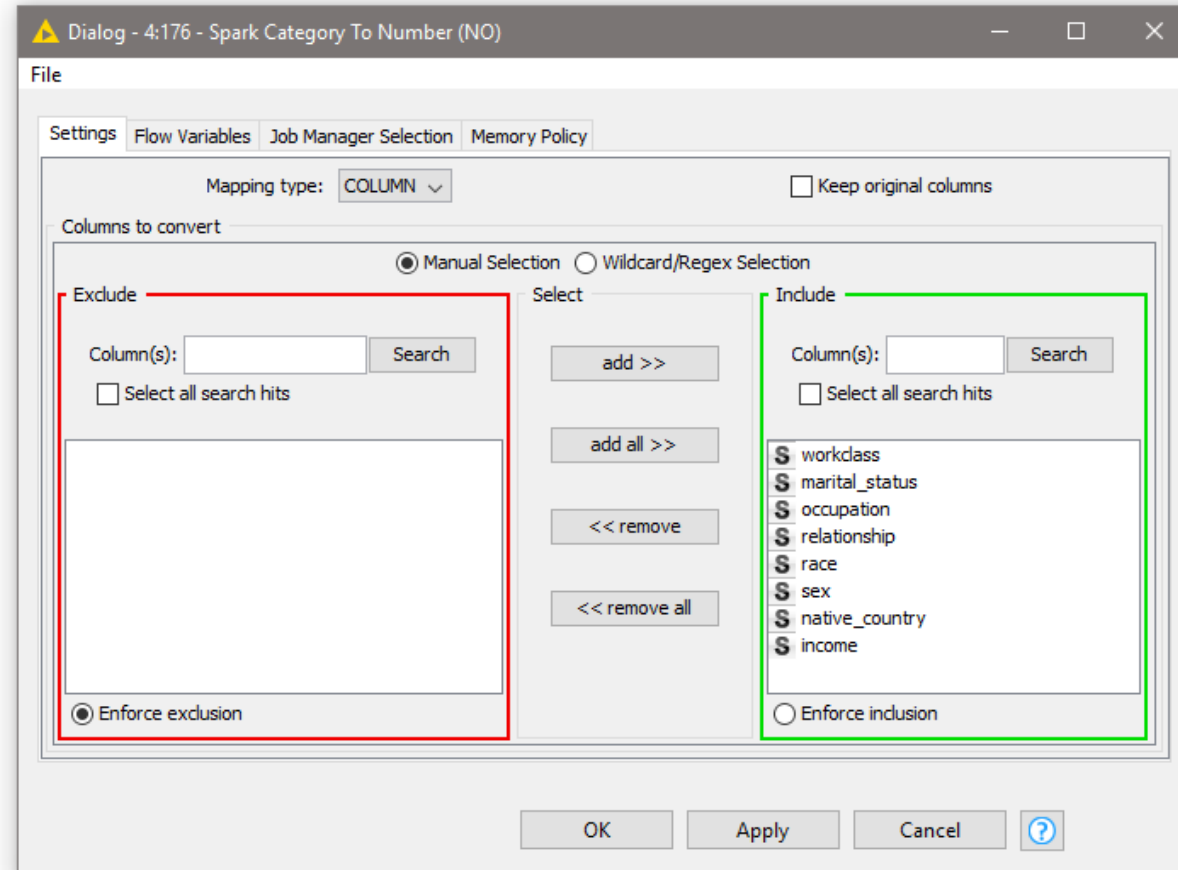
- Import the ss13pme data from Hive into Spark
 - Spark Job URL: *see handout*
 - No authentication required
- Import ss13pme data from the HDFS /input/ss13pme/ folder into Spark

Pre-processing with Spark

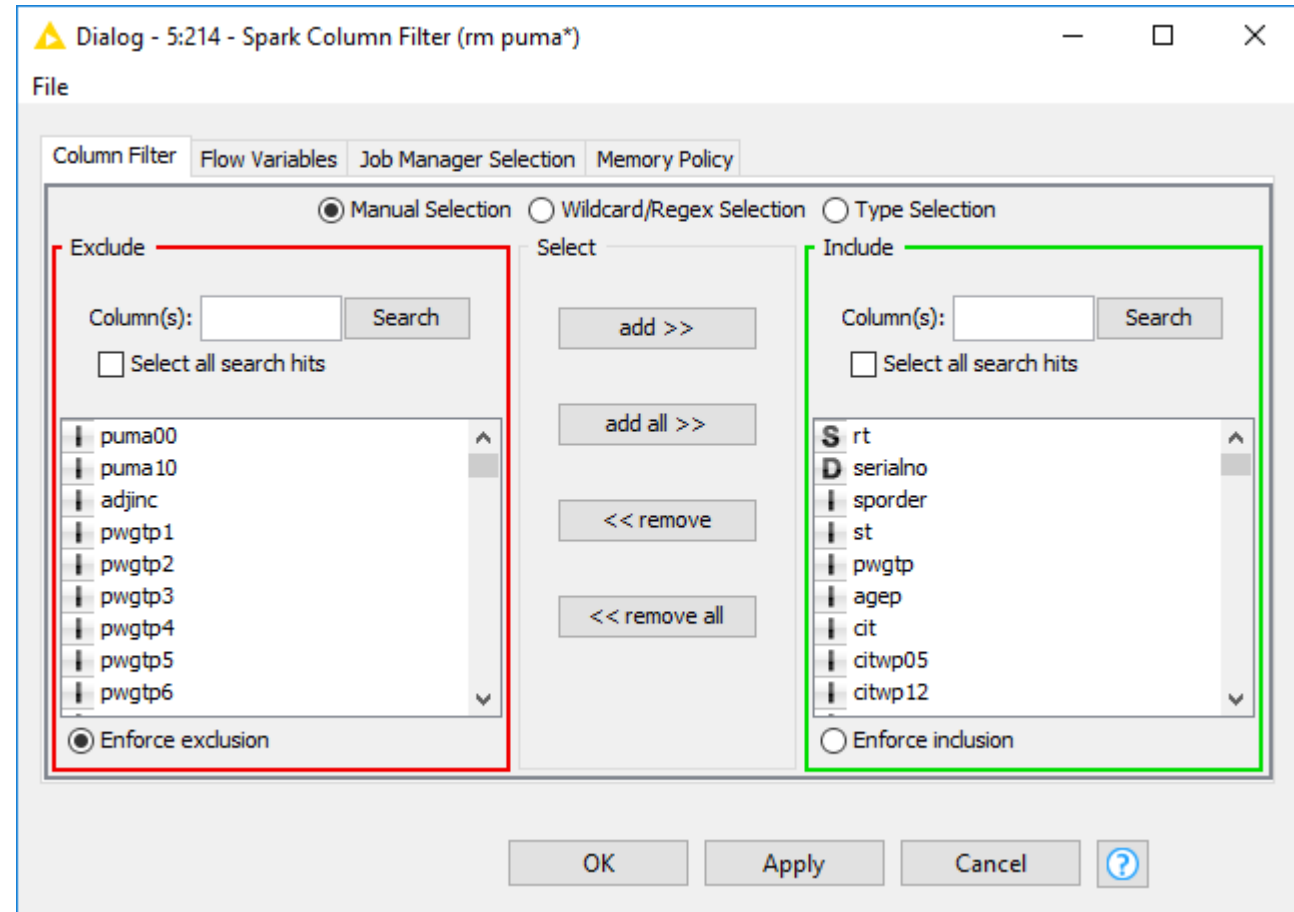
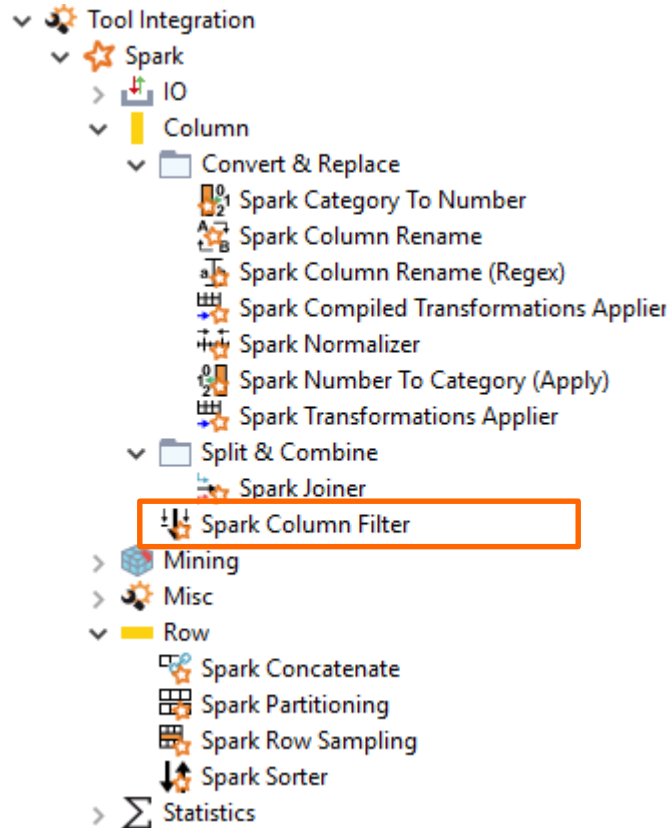
Spark Category to Number



MLlib algorithms only support numeric features and labels

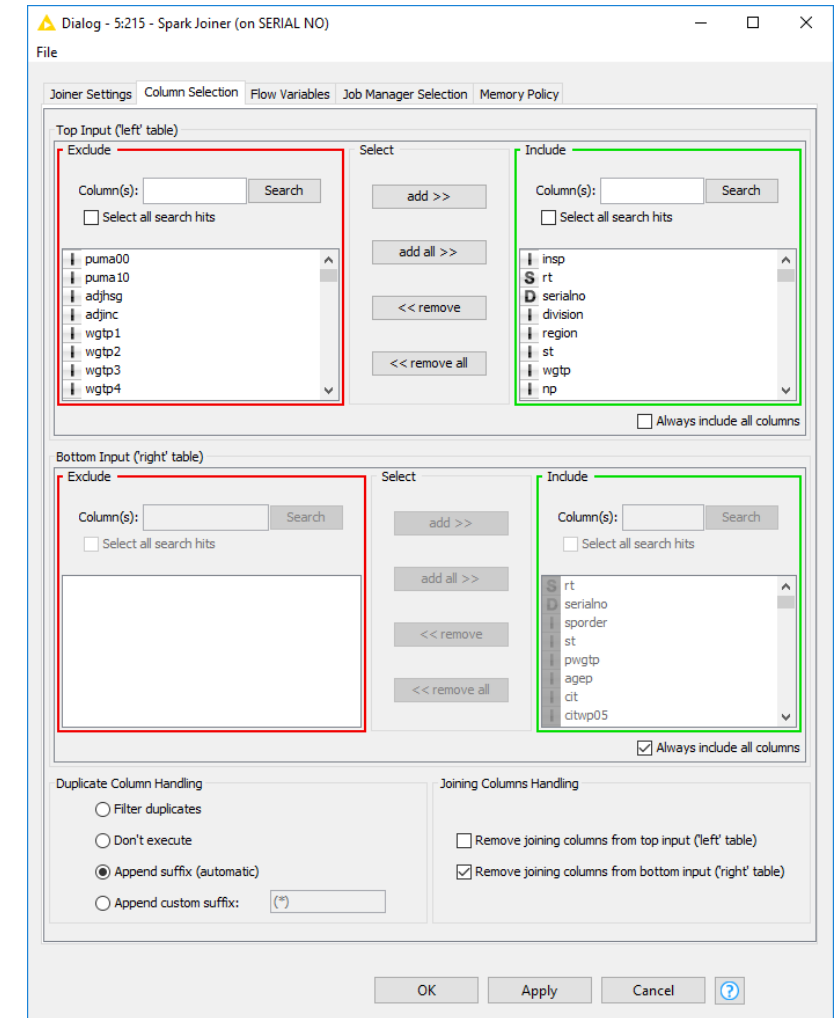
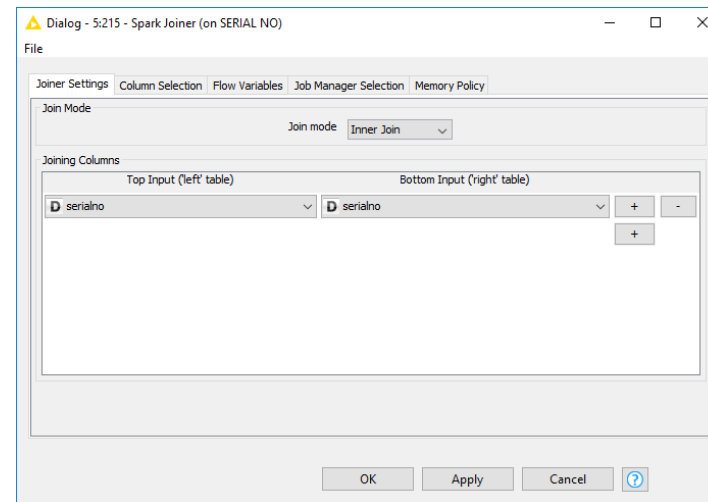


Spark Column Filter

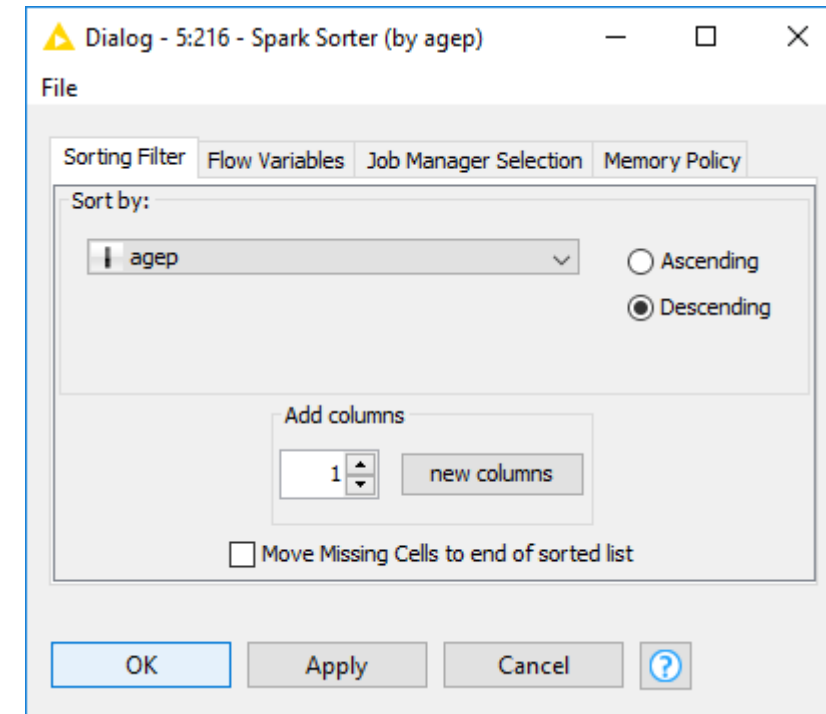
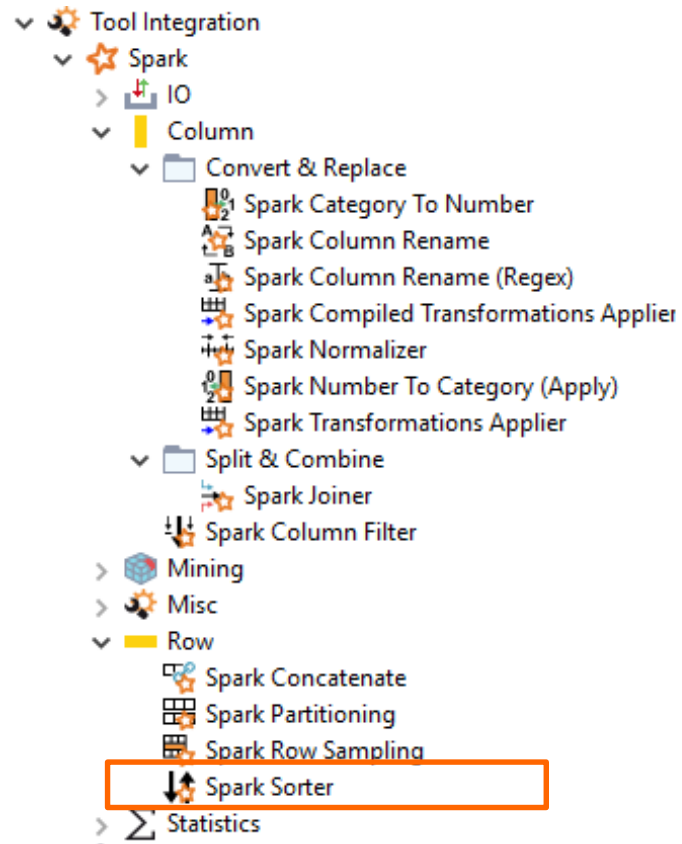


Spark Joiner

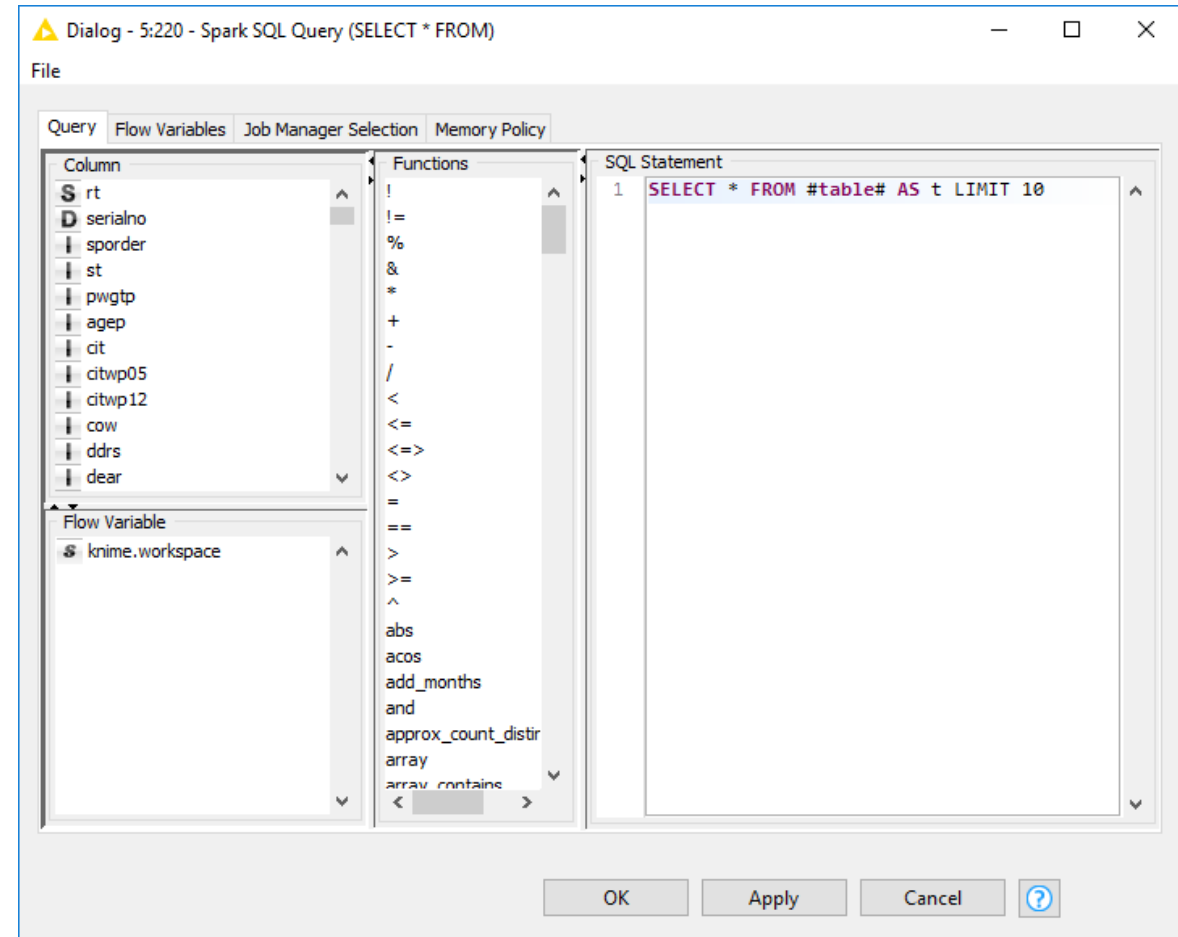
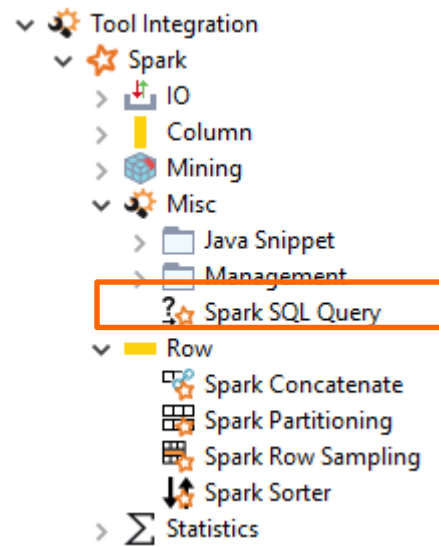
- Tool Integration
 - Spark
 - IO
 - Column
 - Convert & Replace
 - Spark Category To Number
 - Spark Column Rename
 - Spark Column Rename (Regex)
 - Spark Compiled Transformations Applier
 - Spark Normalizer
 - Spark Number To Category (Apply)
 - Spark Transformations Applier
 - Split & Combine**
 - Spark Joiner**
 - Spark Column Filter
 - Mining
 - Misc
 - Row
 - Spark Concatenate
 - Spark Partitioning
 - Spark Row Sampling
 - Spark Sorter
 - Statistics



Spark Sorter

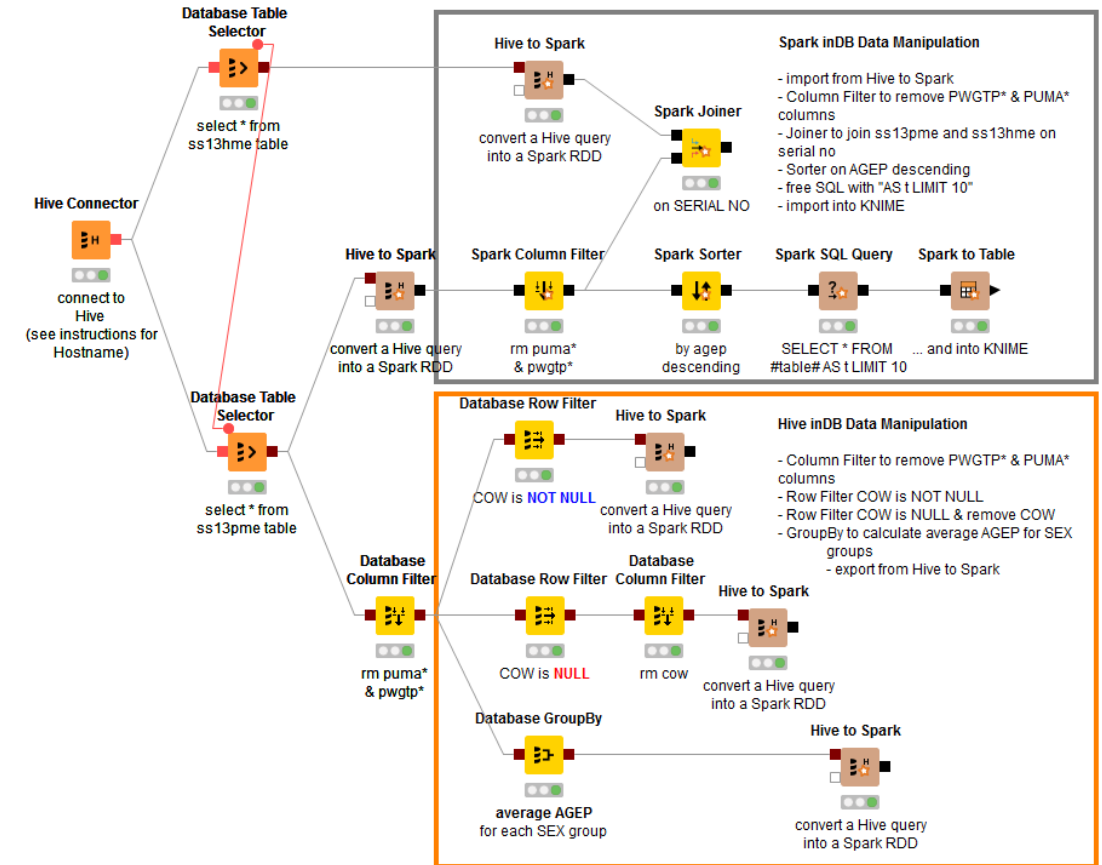


Spark SQL Query



Mix & Match

Thanks to the transferring nodes (Hive to Spark and Spark to Hive, Table to Spark and Spark to Table) you can mix and match in-database processing operations



Section Exercise – 02_Spark_InDB_Processing

This workflow mixes **Hive in DB Manipulation with Spark in DB manipulation.**

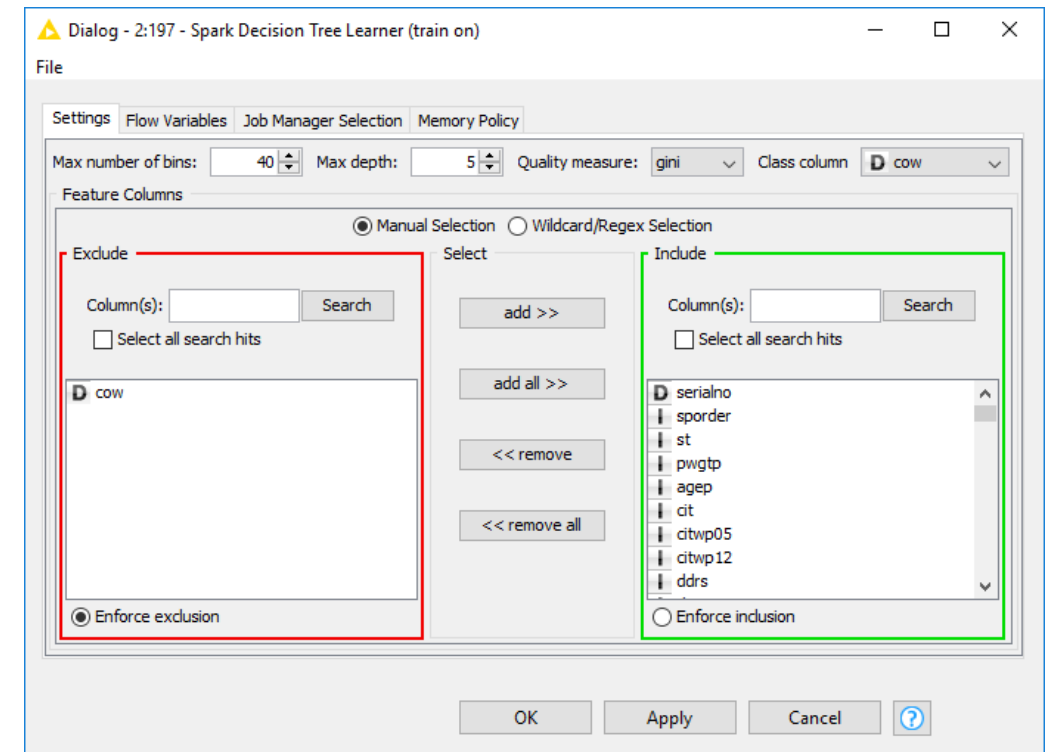
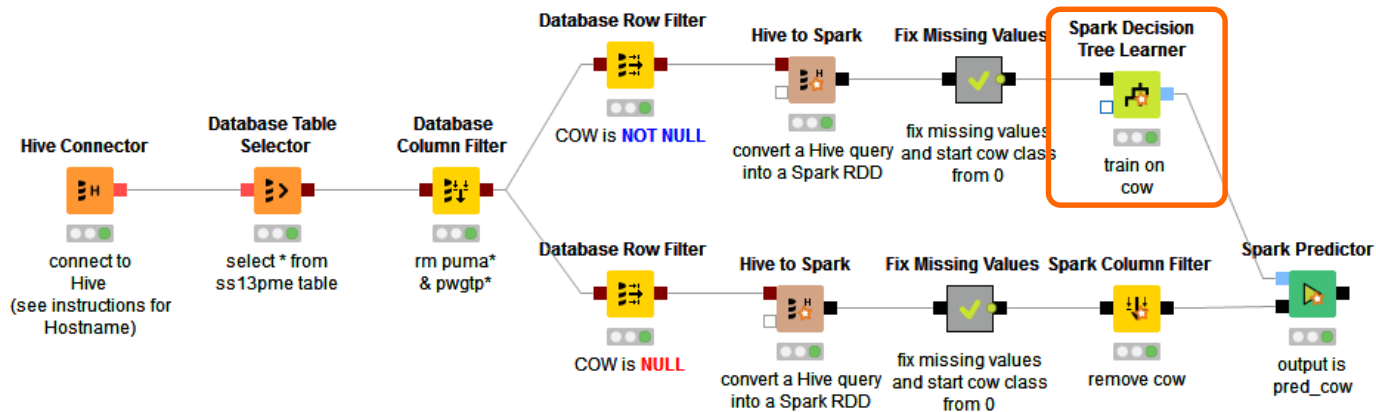
Hive in DB is already present with Database Column Filter, Row Filter, and GroupBy nodes followed by Hive to Spark nodes. Use the following Spark in DB processing nodes:

- Column Filter to remove PWGTP* and PUMA* columns
- Joiner to join ss13pme and ss13hme tables on SERIAL NO
- Sorter to sort on AGE descending
- Use free SQL code to extract top 10 data rows
- import results into KNIME

Machine Learning with Spark

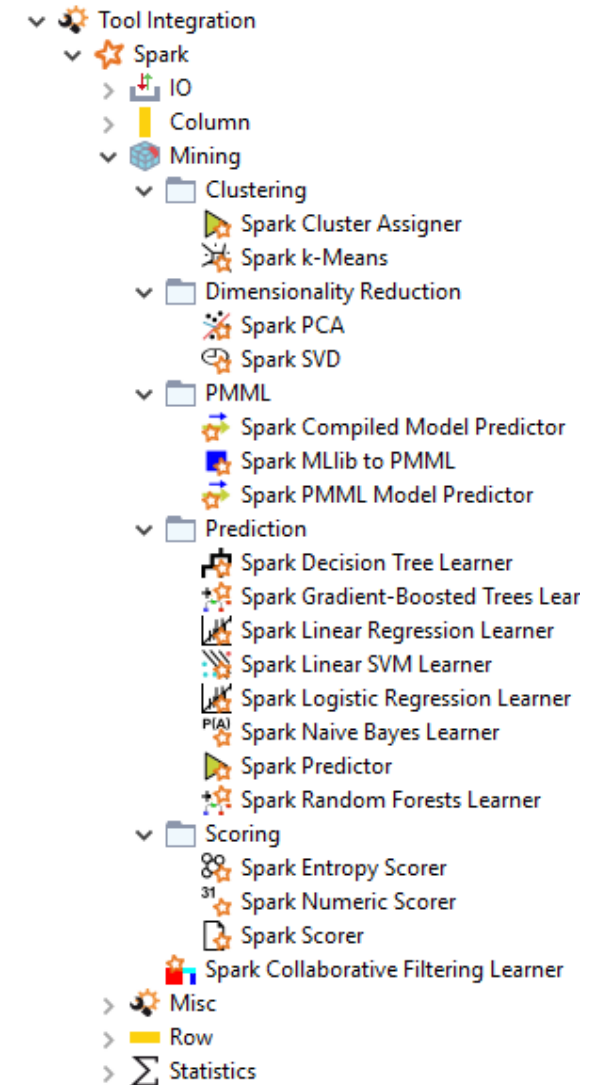
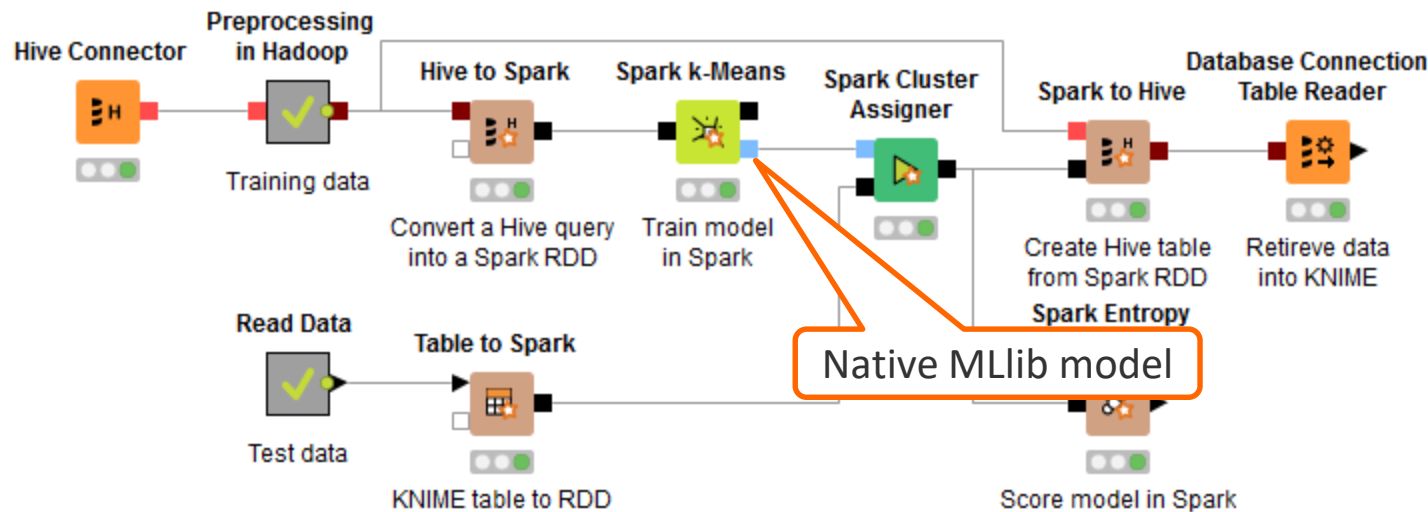
MLlib Integration: Spark Decision Tree

- Usage model and dialogs similar to existing nodes
- No coding required



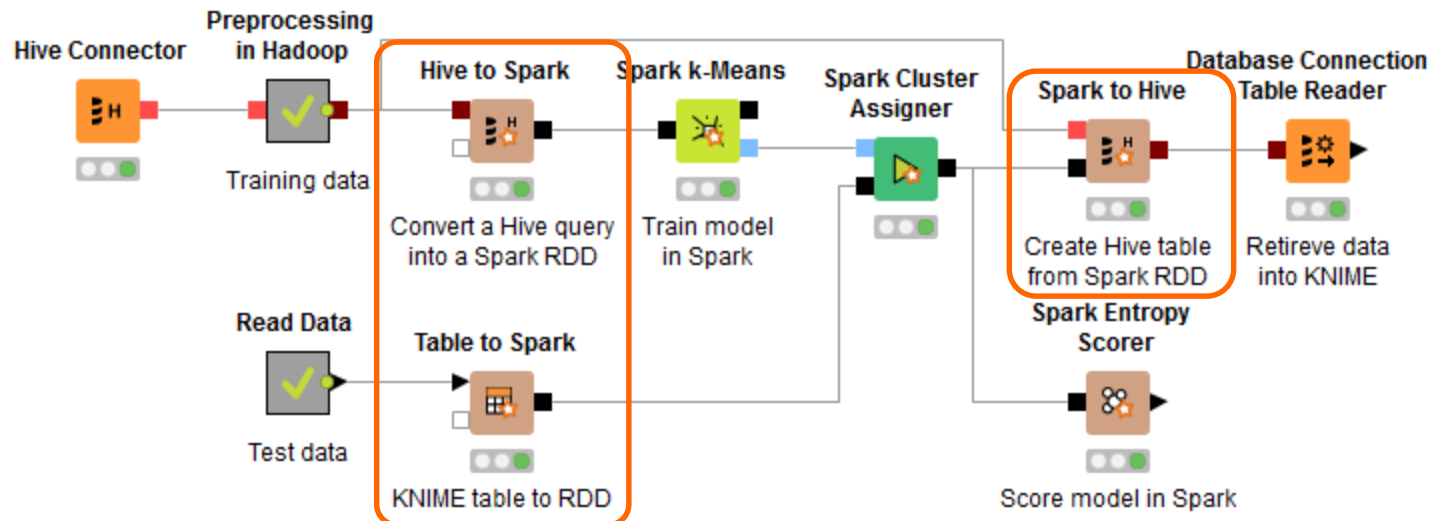
MLlib Integration: Spark k-Means

- MLlib model ports for model transfer
- Native MLlib model learning and prediction
- Spark nodes start and manage Spark jobs
- Supports Spark job cancelation



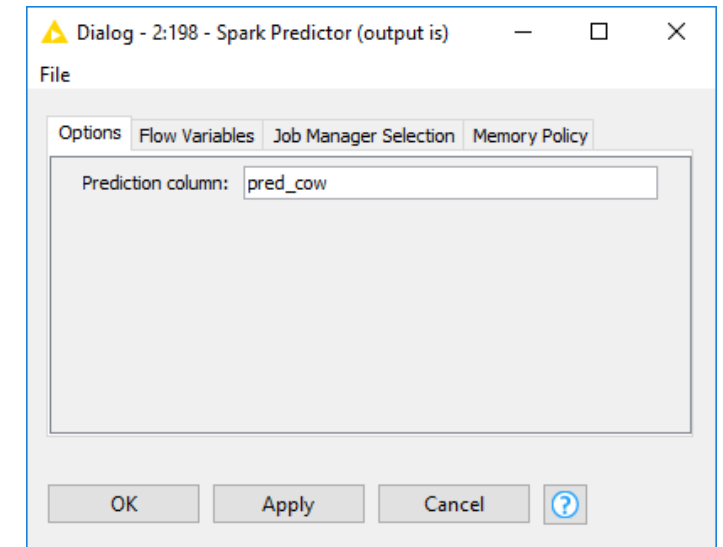
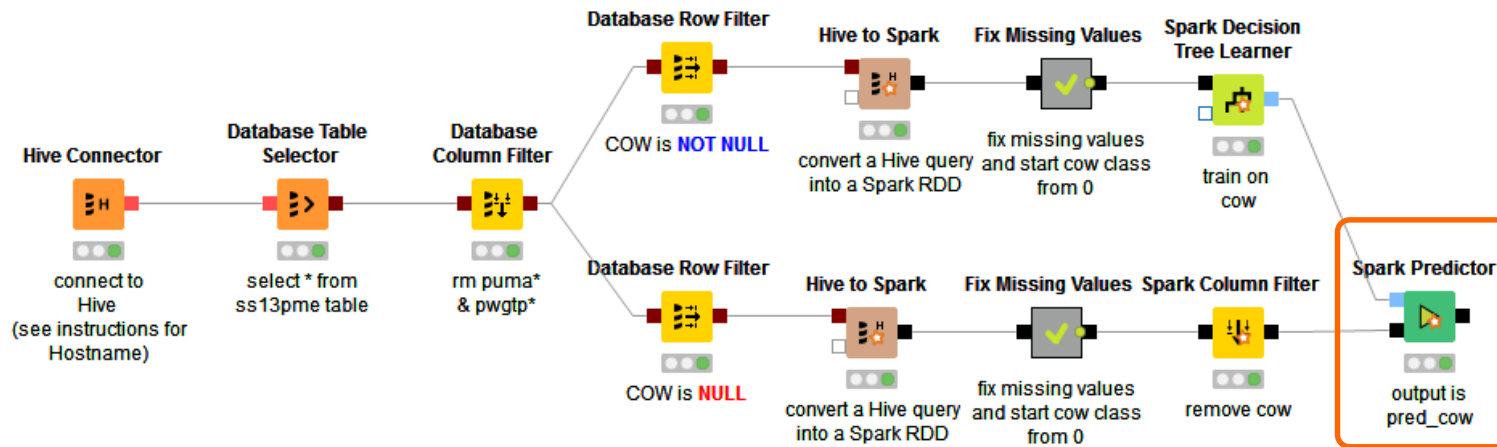
MLlib Integration stays in Spark

- Spark RDDs as input/output format
- Data stays within your cluster
- No unnecessary data movements
- Several input/output nodes e.g. Hive, hdfs files, ...



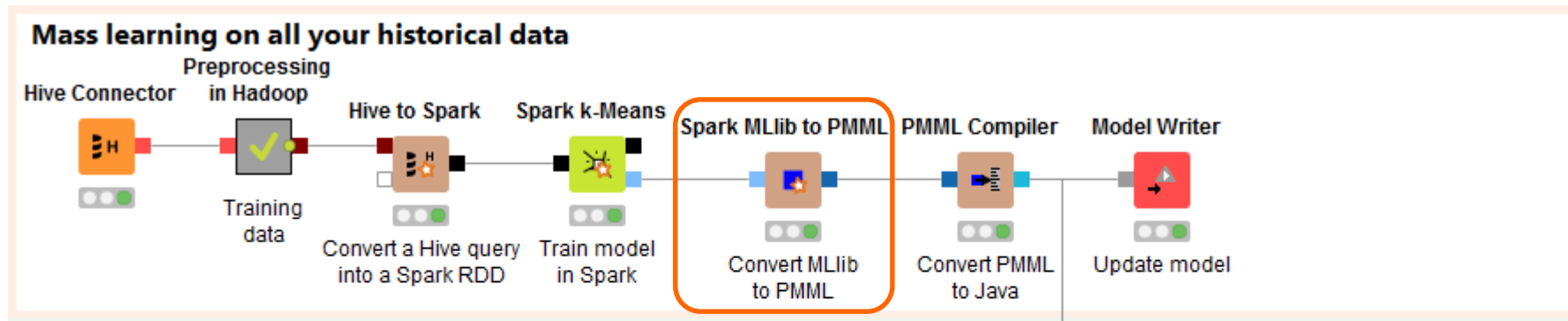
Mllib Integration: Spark Predictor

- Algorithms only support numeric features and labels
- Tree algorithms have optional category mapping input port
- Spark Predictor assigns labels based on a given supervised model



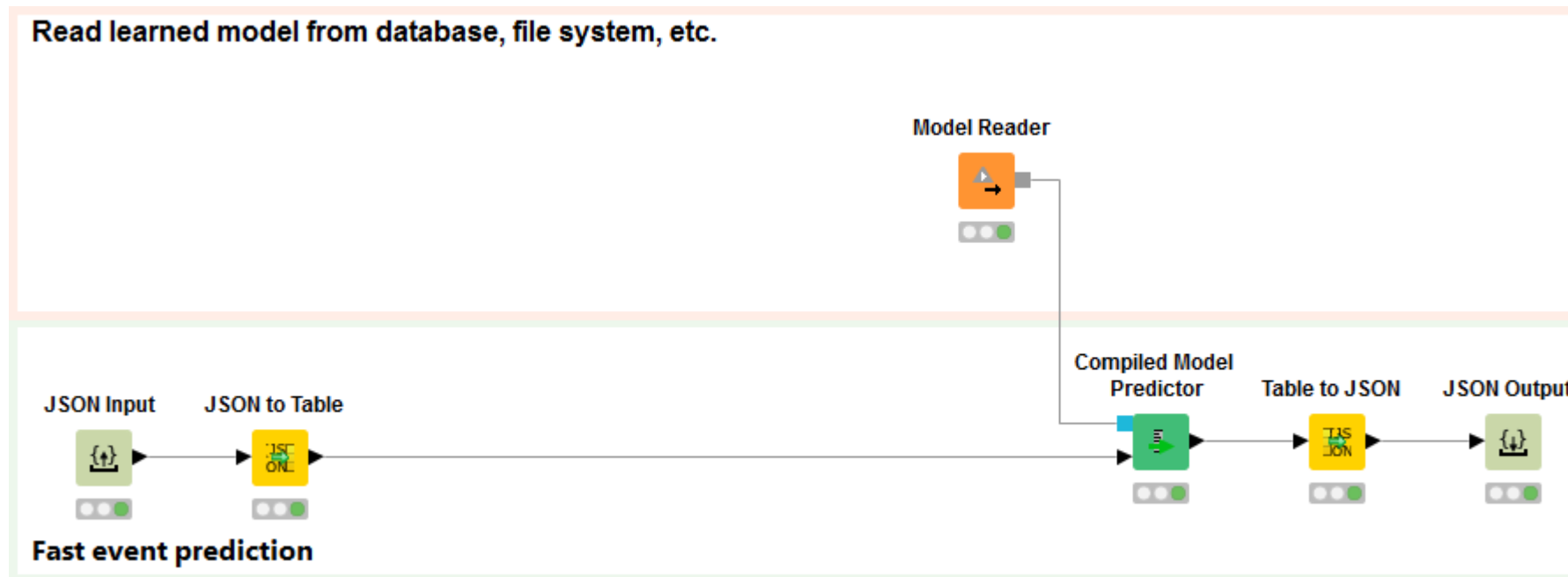
Mass Learning in Spark – Conversion to PMML

- Mass learning on Hadoop
- Convert supported MLlib models to PMML



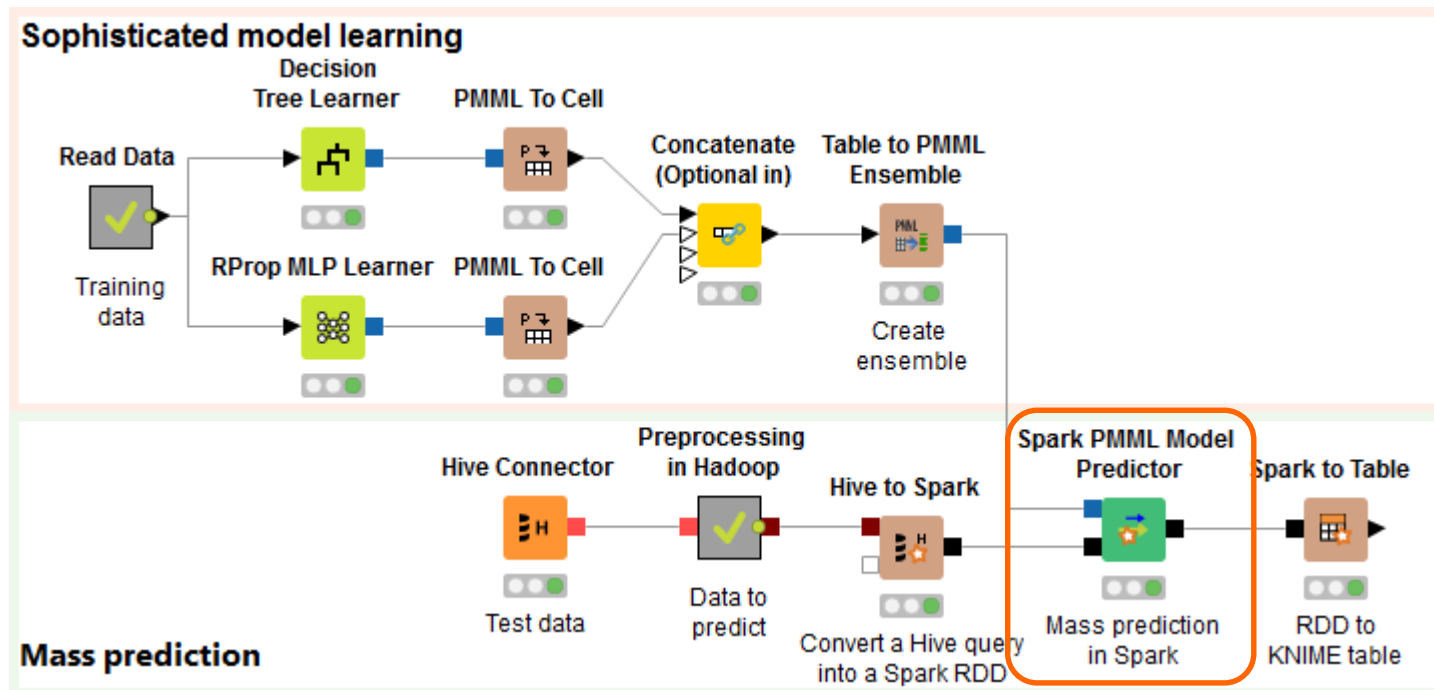
Mass Learning in Spark – Fast Event Prediction in KNIME on Demand

- Fast event prediction based on compiled models

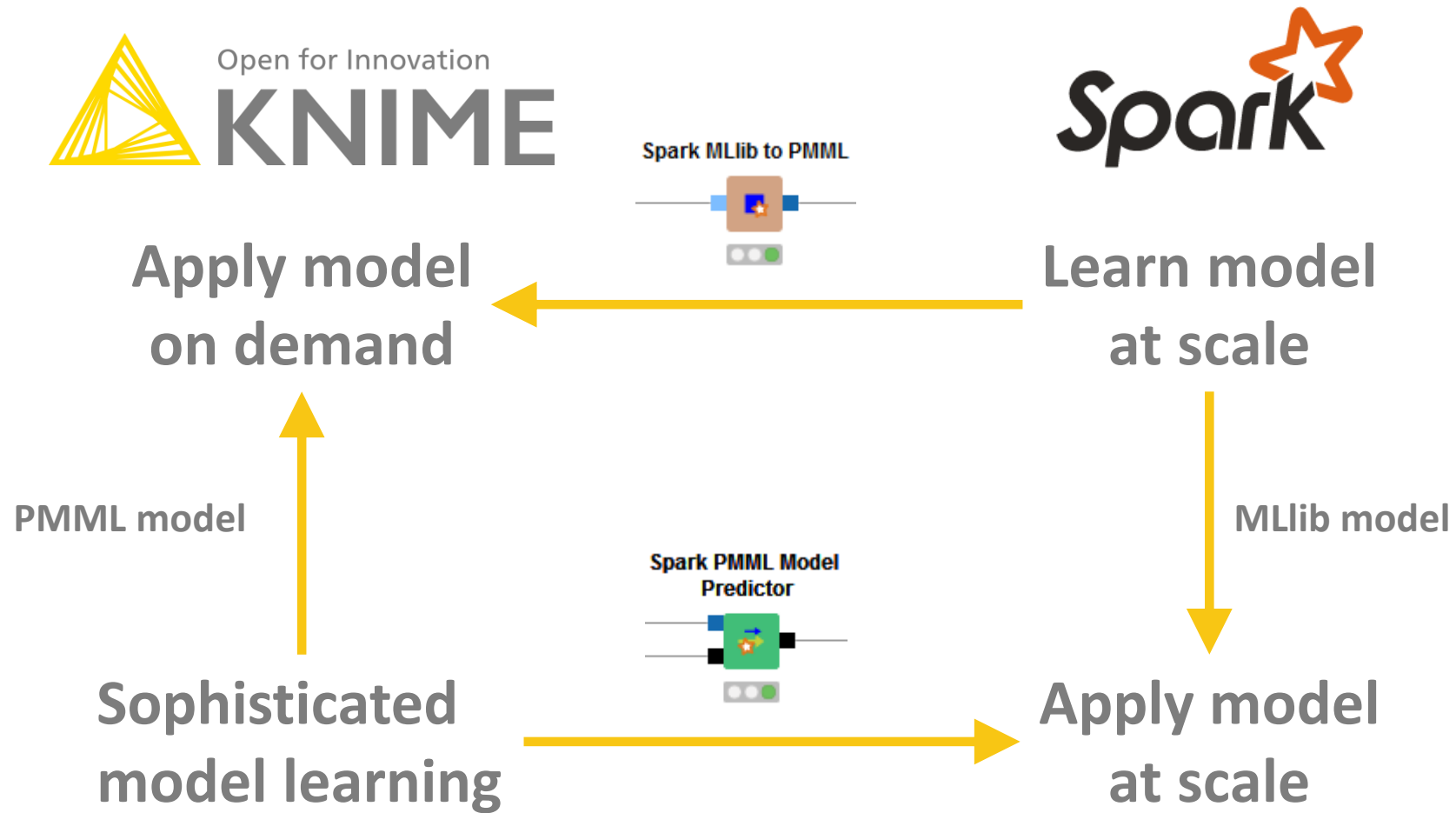


Sophisticated Learning in KNIME - Mass Prediction in Spark

- Supports KNIME models and pre-processing steps
- Sophisticated model learning in KNIME
- Mass prediction on Hadoop



Closing the Loop



Section Exercise – 03_Spark_Modelling

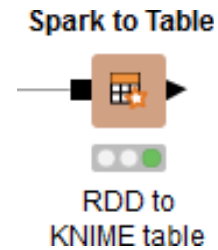
On the ss13pme table, the current workflow isolates the rows with not missing COW value from data rows with missing COW value, fixes missing values, and for the last subset removes the COW column.

- Train a decision tree on COW on data rows where COW is NOT NULL
- Apply the decision tree model to predict COW value on rows with missing COW

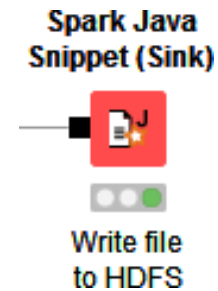
Export Data back into KNIME/Hadoop

Export Data to KNIME/Hadoop

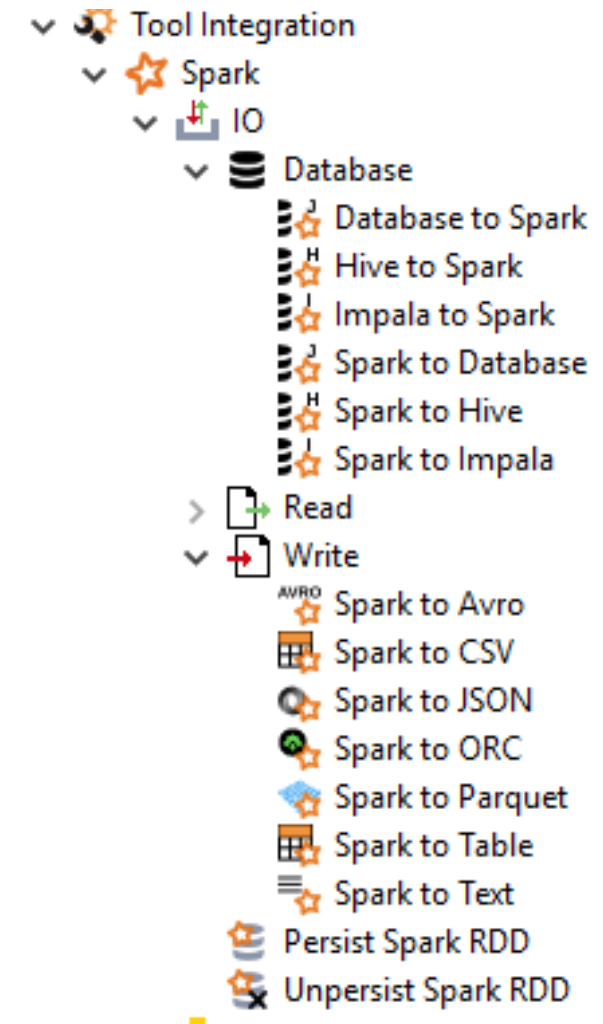
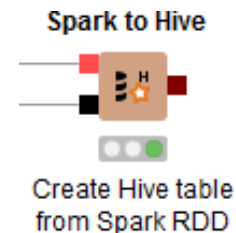
- To KNIME



- Write to HDFS



- To Hive



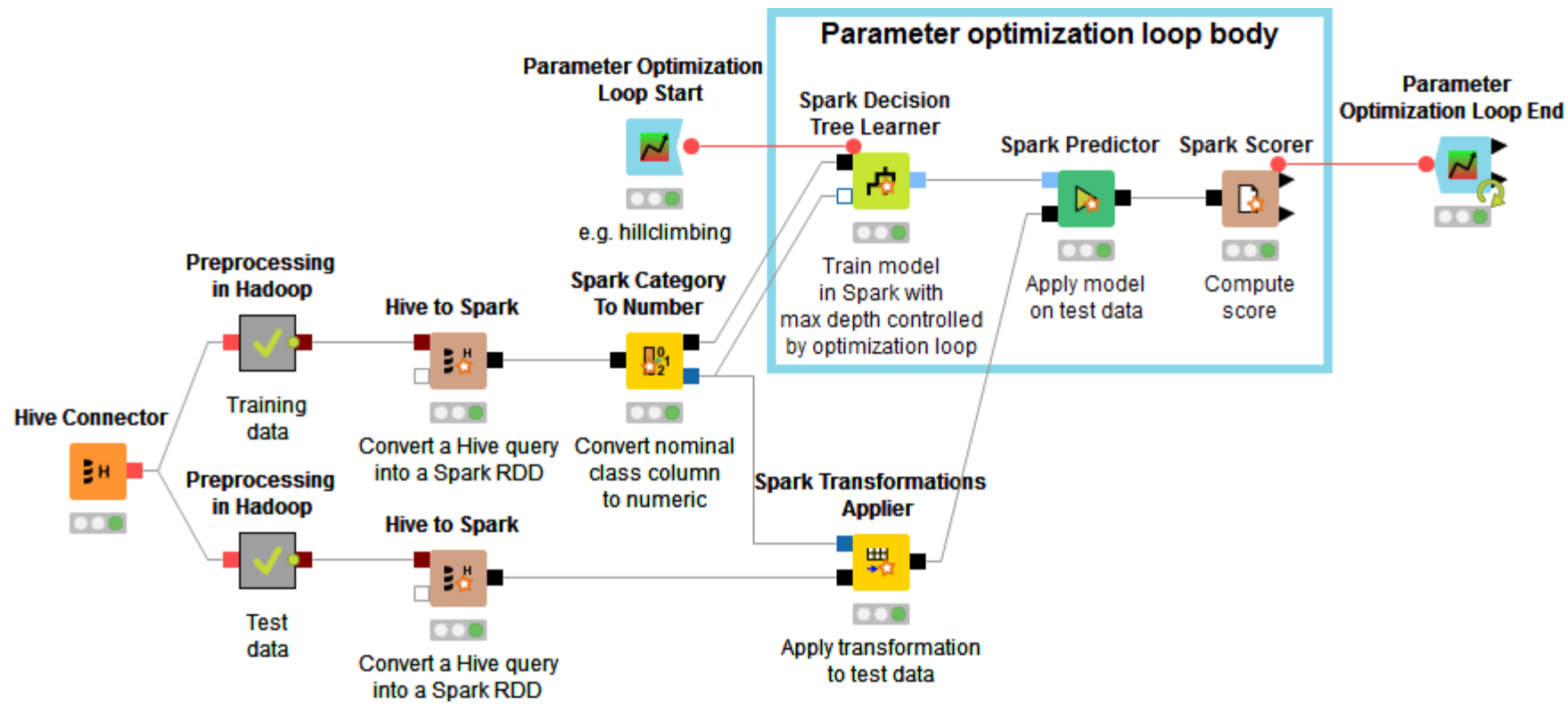
Section Exercise – 04_Spark_WritingToDB

This workflow implements a Spark predictor to predict COW values from the ss13pme data set. The model is applied to predict COW values where they are missing. Now export the new data set without missing values to:

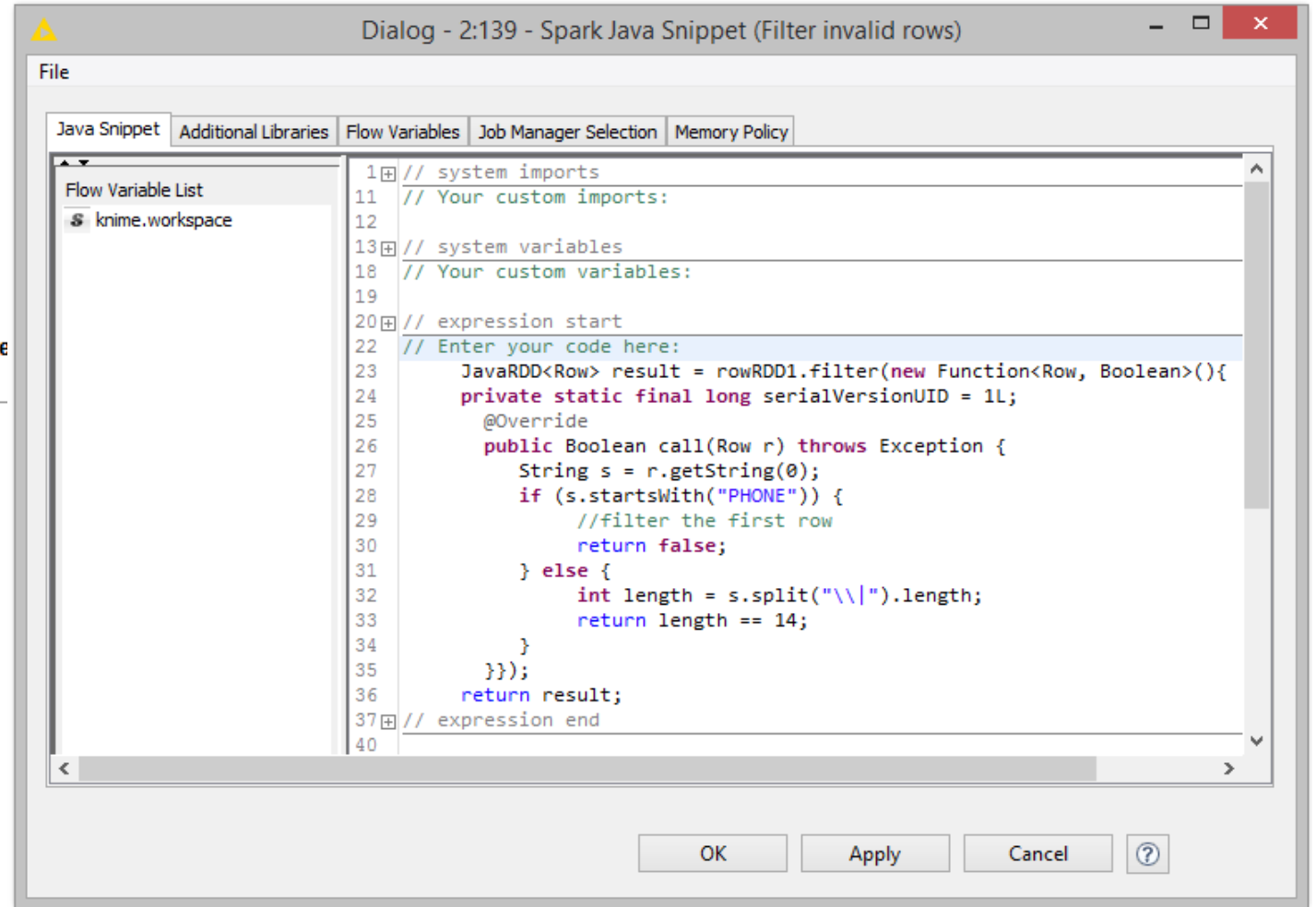
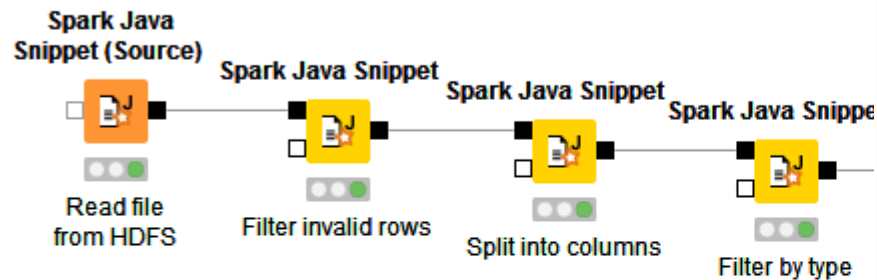
- KNIME table,
- parquet on Spark,
- Hive.

Mix and Match

- KNIME <-> Hive <-> Spark

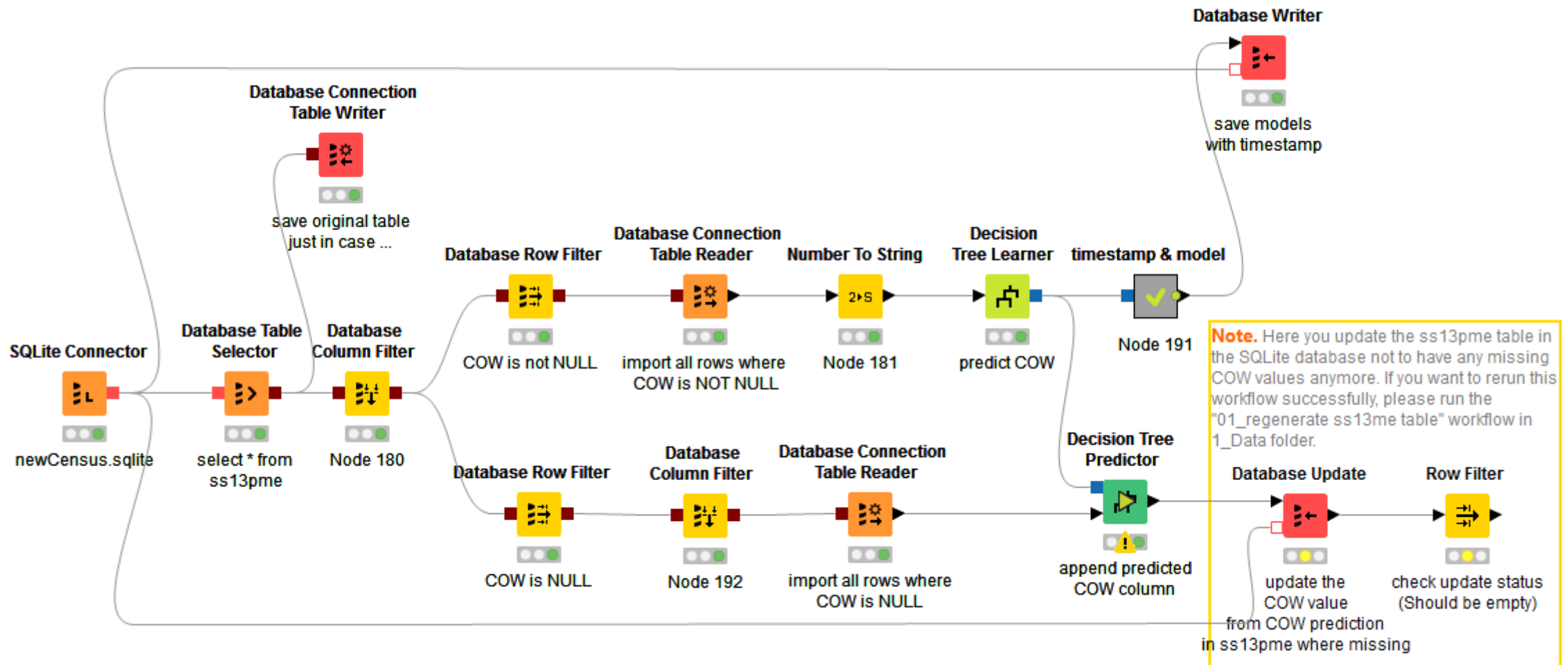


Modularize and Execute Your Own Spark Code

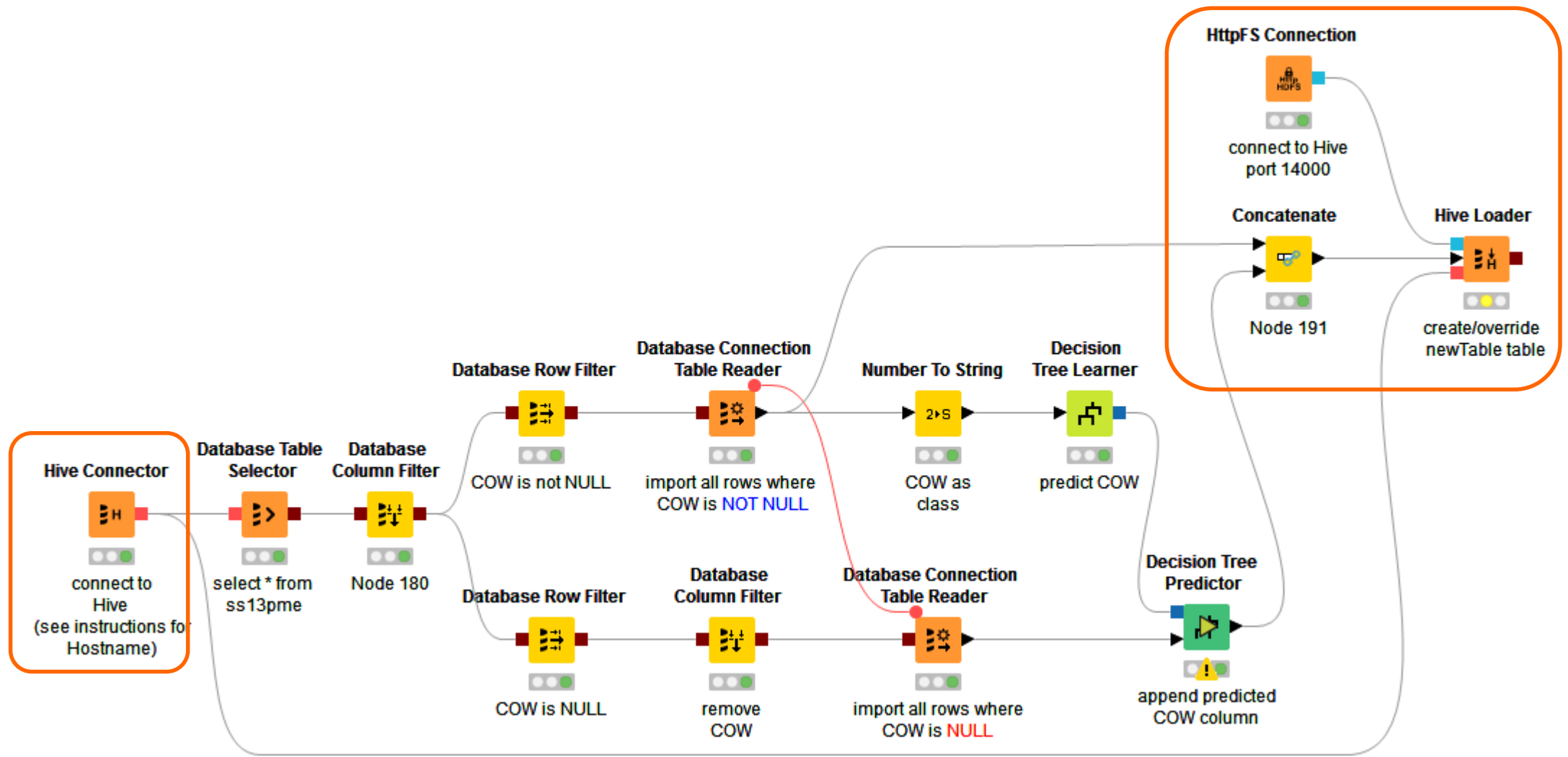


Conclusions

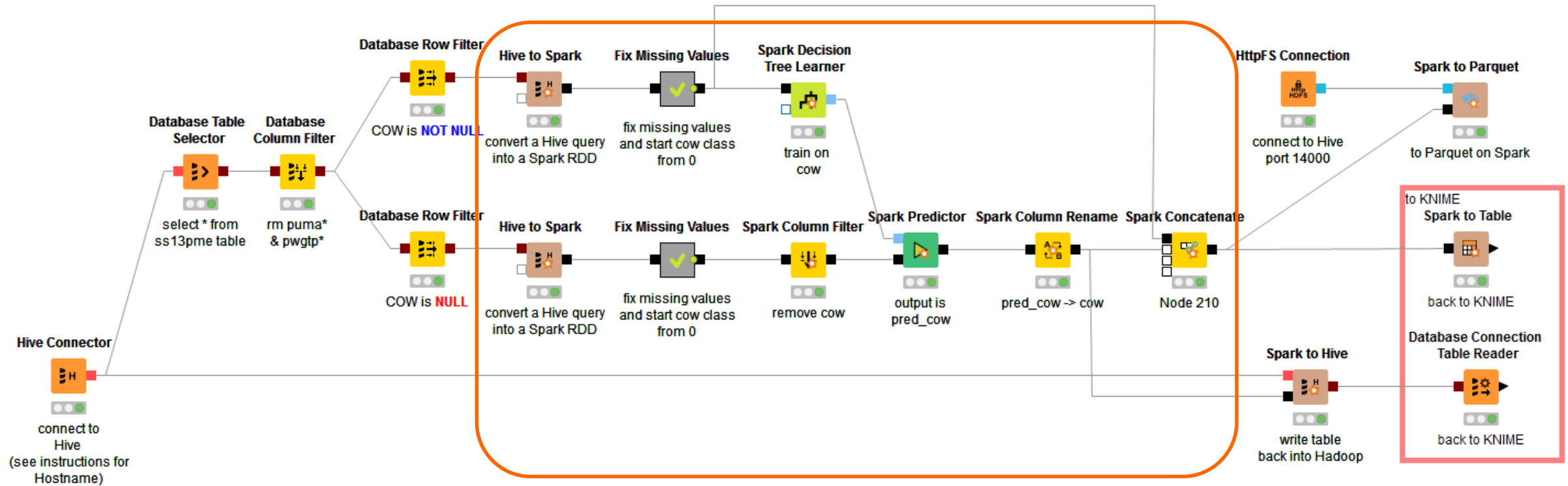
SQLite



Hadoop Hive



Spark



Want to try it at home?

- Hadoop cluster
 - Use your own Hadoop cluster
 - Use a preconfigured virtual machine
 - <http://hortonworks.com/products/hortonworks-sandbox/>
 - http://www.cloudera.com/downloads/quickstart_vms.html
- Download and install compatible Spark Job Server
 - See installation steps at <https://www.knime.org/knime-spark-executor#install>
- **For a free 30-day Trial go to**
<https://www.knime.org/knime-big-data-extensions-free-30-day-trial>



The End

education@knime.com

Twitter: @KNIME

