



Open for Innovation

**KNIME**

# Welcome to KNIME Big Data Workshop

Going live at:

Chicago 10:00 am

San Francisco 8:00 am

New York 11:00 am

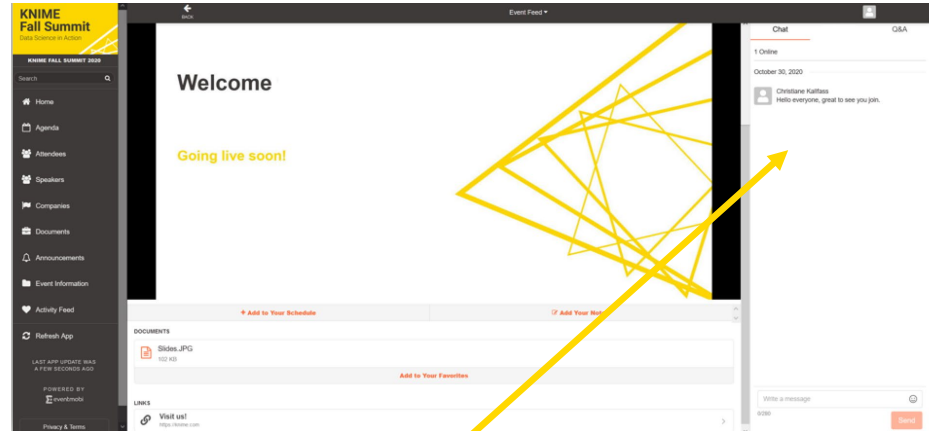
Berlin 5:00 pm



# Housekeeping

- Post in the chat where you are dialing and discuss with other attendees
- Questions? Post them in the Q&A

Questions will be answered after the presentation.



# What is "Big Data" about?

---

1

*"...ways to analyze, systematically extract information from [...] data sets that are too large or complex to be dealt with by traditional data-processing application software."* [1]

2

The three Vs of what makes data "big":

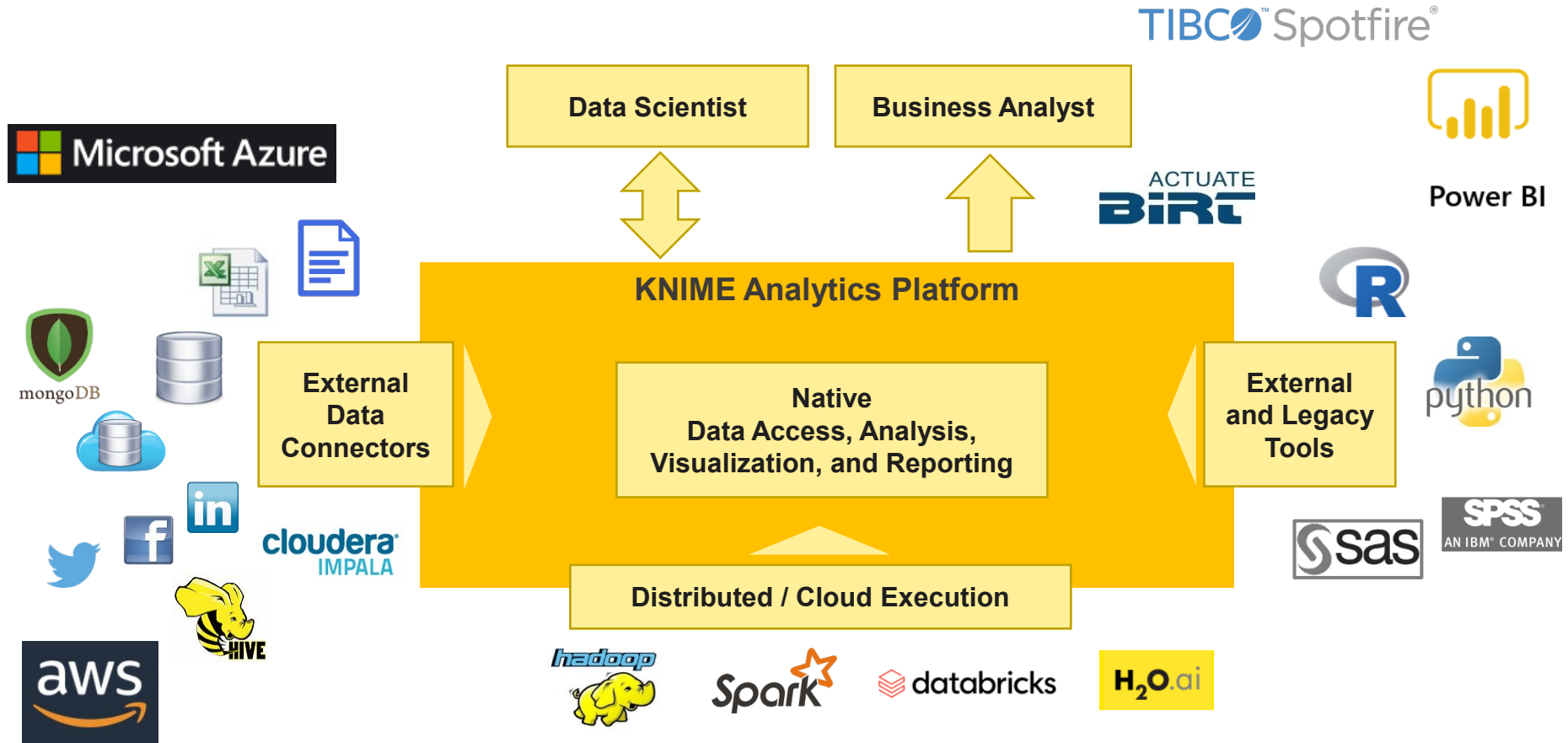
- Volume (size of data)
- Variety (tabular, text, images, video, audio, time series, ...)
- Velocity (produced fast and continuously)

**Goal of big data technologies:**

Enable predictive or other types of advanced analytics to extract **value** from big data.

[1] [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data)

# KNIME Analytics Platform: Open for Every Data, Tool, and User



# Agenda

---



Introduction to Hadoop and Spark



KNIME Big Data Connectors



KNIME Extension for Apache Spark



KNIME H2O Sparkling Water Integration



KNIME Workflow Executor for Apache Spark

# Apache Hadoop

---



Open-source project for distributed storage and processing of large data sets



Designed to scale up to thousands of machines



First release in 2006

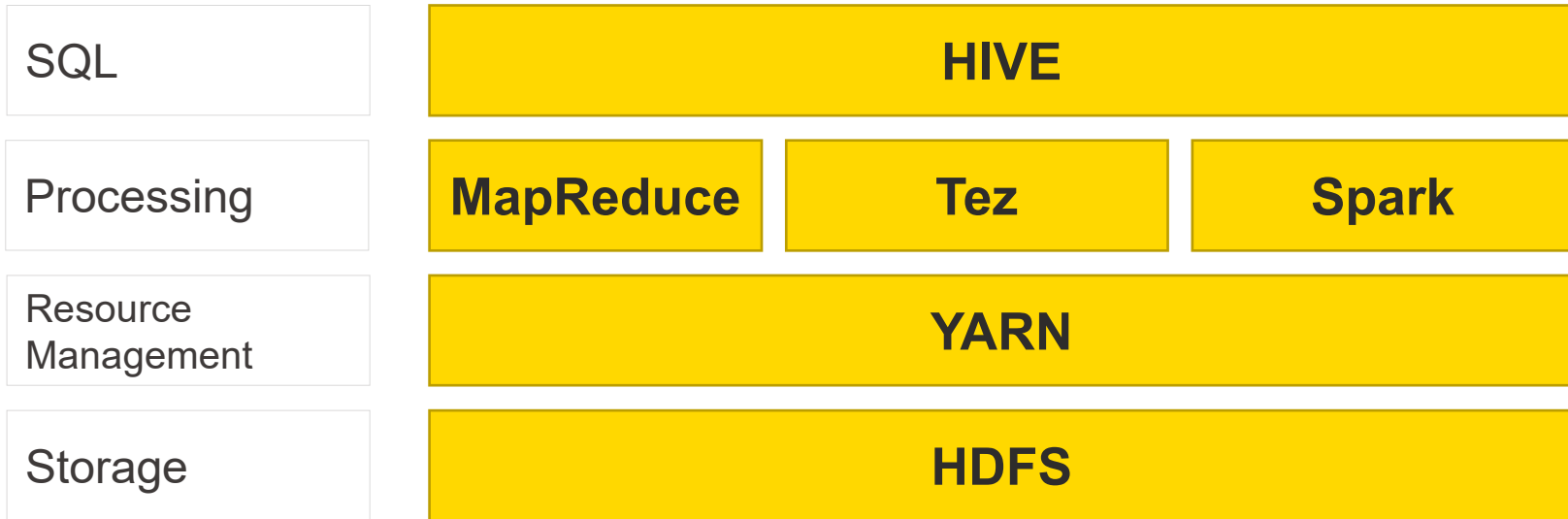
Rapid adoption, promoted to top level Apache project in 2008  
Inspired by Google File System (2003) paper



Spawned diverse ecosystem of products

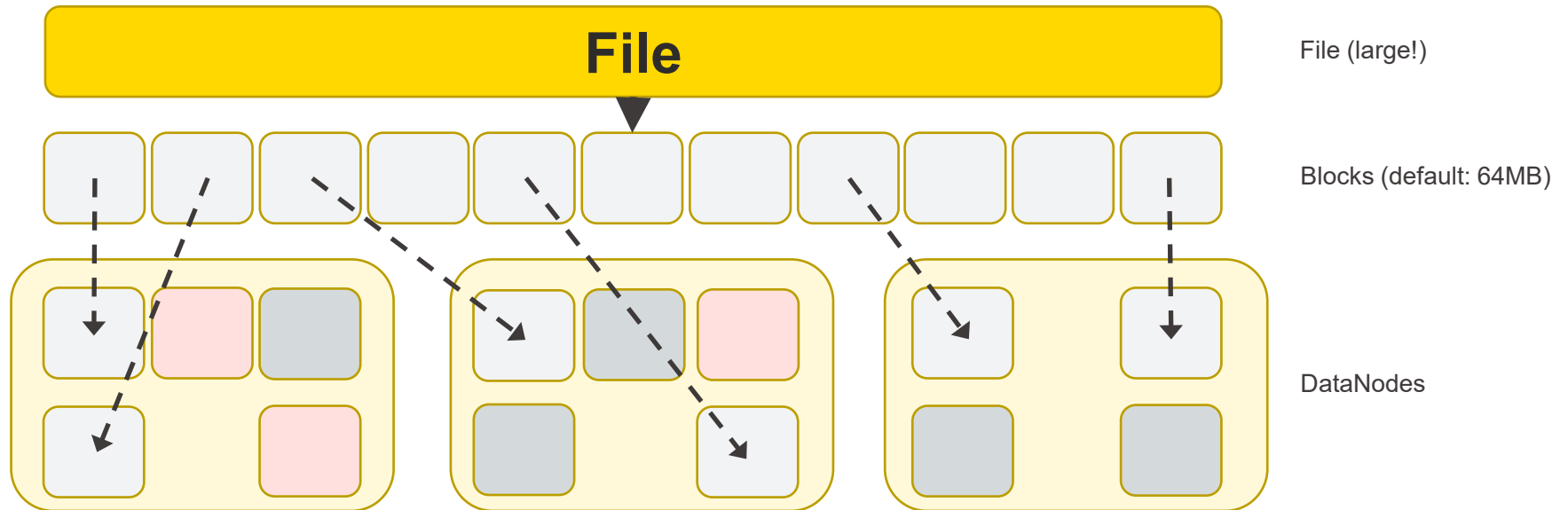
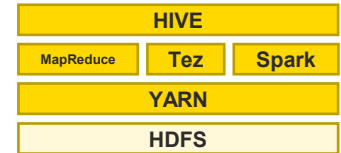
# Hadoop Ecosystem

---



# HDFS

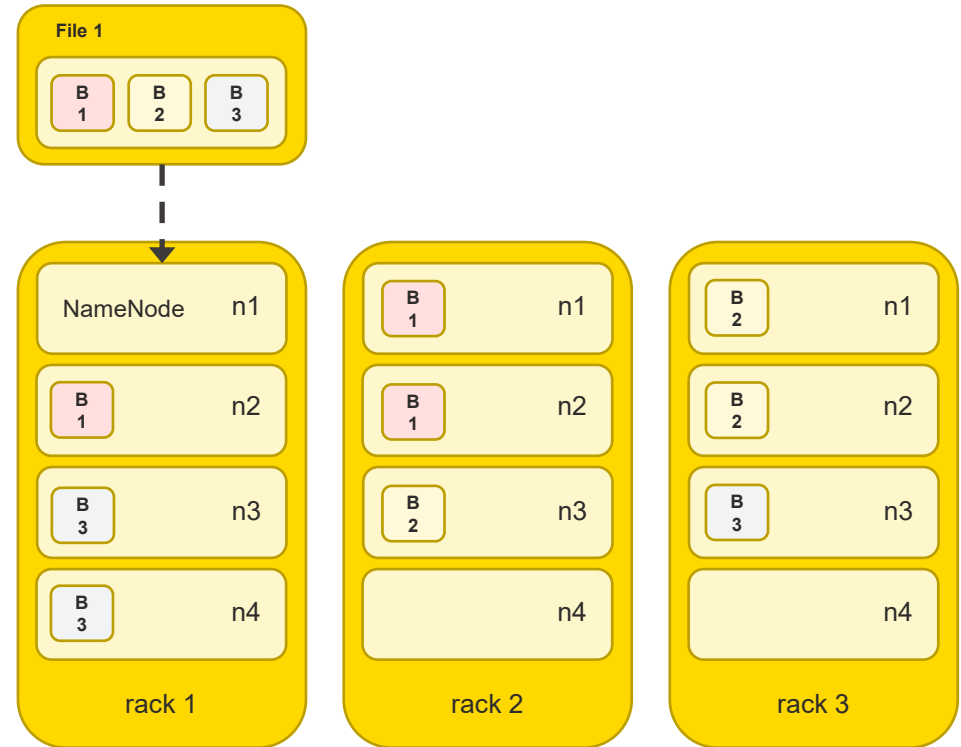
- Hadoop distributed file system
- Stores large files across multiple machines





# HDFS – Data Replication and File Size

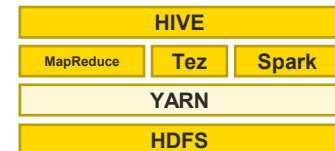
- Data Replication
- All blocks of a file are stored as sequence of blocks
- Blocks of a file are replicated for fault tolerance (usually 3 replicas)
  - improves data reliability, availability, and network bandwidth utilization



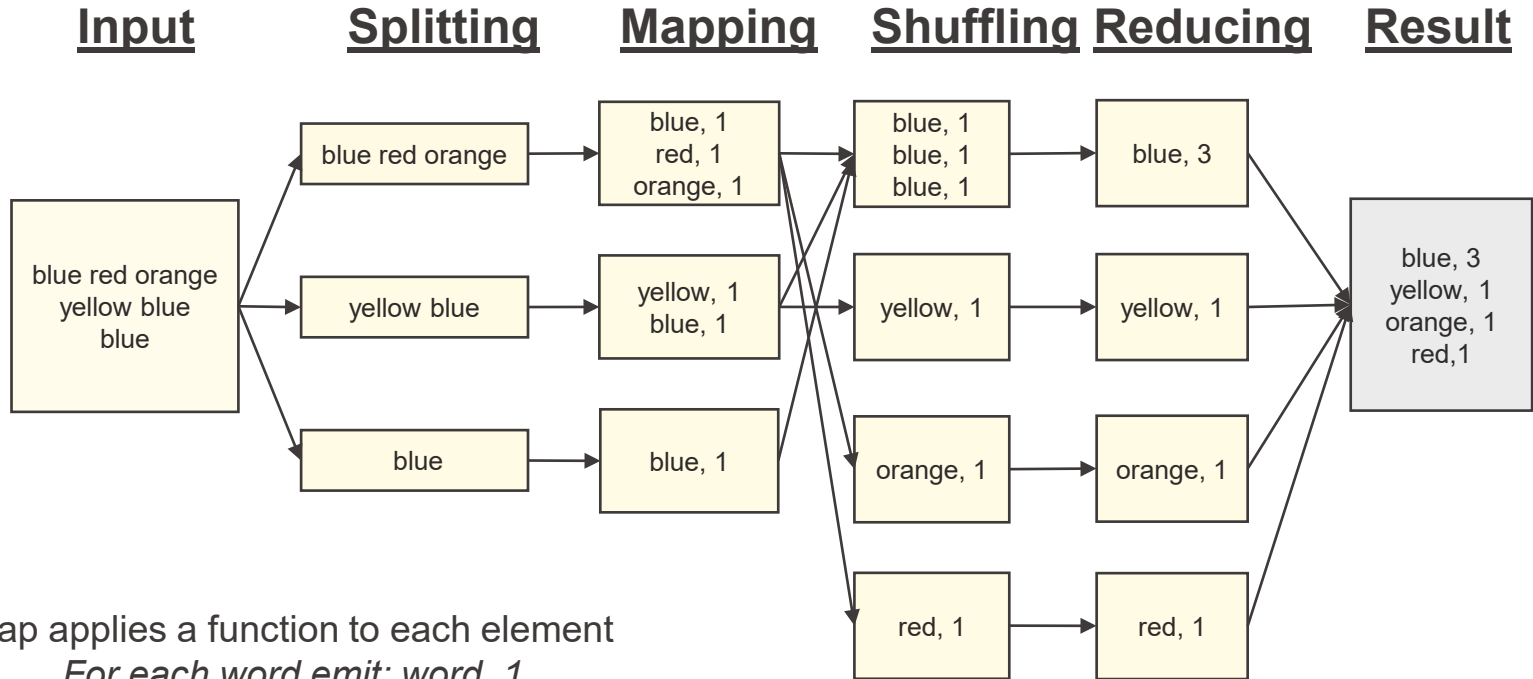
# YARN

---

- Cluster resource management system
- Two elements
  - Resource manager (one per cluster):
    - Knows where worker nodes 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



# MapReduce

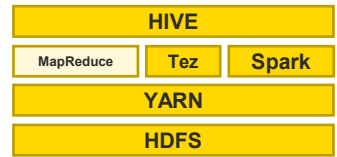


Map applies a function to each element

*For each word emit: word, 1*

Reduce aggregates a list of values to one result

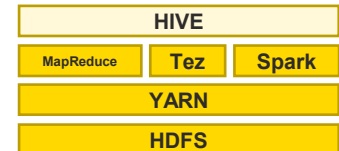
*For all equal words sum up count*



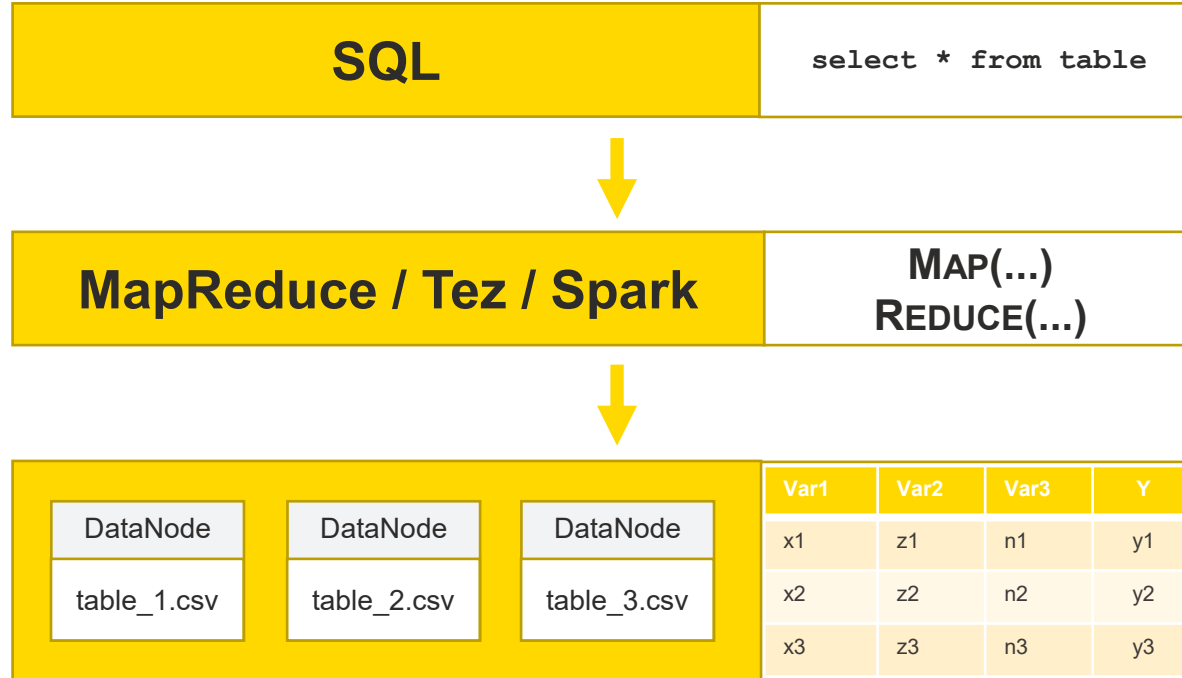
# Hive

---

- SQL database on top of files in HDFS
- Provides data summarization, query, and analysis
- Interprets a set of files as a database table (schema information to be provided)
- Translates SQL queries to MapReduce, Tez, or Spark jobs
- Supports various file formats:
  - Text/CSV
  - SequenceFile
  - Avro
  - ORC
  - Parquet



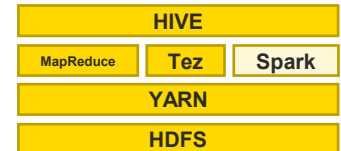
# Hive



# Spark

---

- Cluster computing framework for large-scale data processing
- In-memory computing
  - much (!) faster than MapReduce
- Programmatic interface (Scala, Java, Python, R)
- Great for:
  - Iterative algorithms
  - Interactive data analysis



# Spark – DataFrame

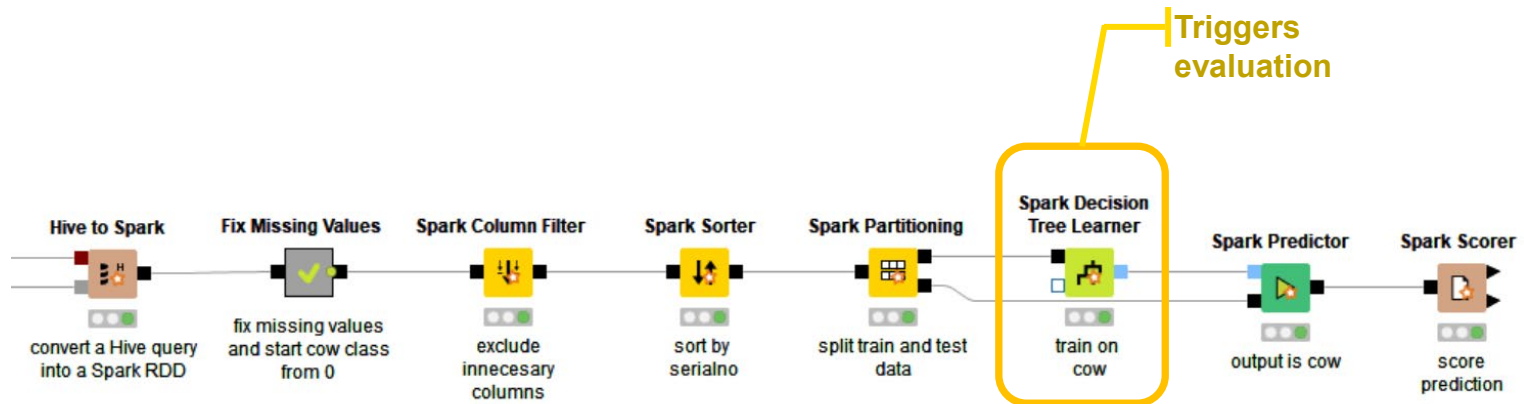
---

- Table-like: Collection of rows, organized in columns with names and types
- Immutable:
  - Data manipulation = creating new DataFrame from an existing one by applying a function on it
- Lazily evaluated:
  - Functions are not executed until an action is triggered, that requests to actually see the row data
- Distributed:
  - Each row belongs to exactly one partition
  - Each partition is held by a Spark Executor

Name	Surname	Age
John	Doe	35
Jane	Roe	29
...	...	...

# Spark – Lazy Evaluation

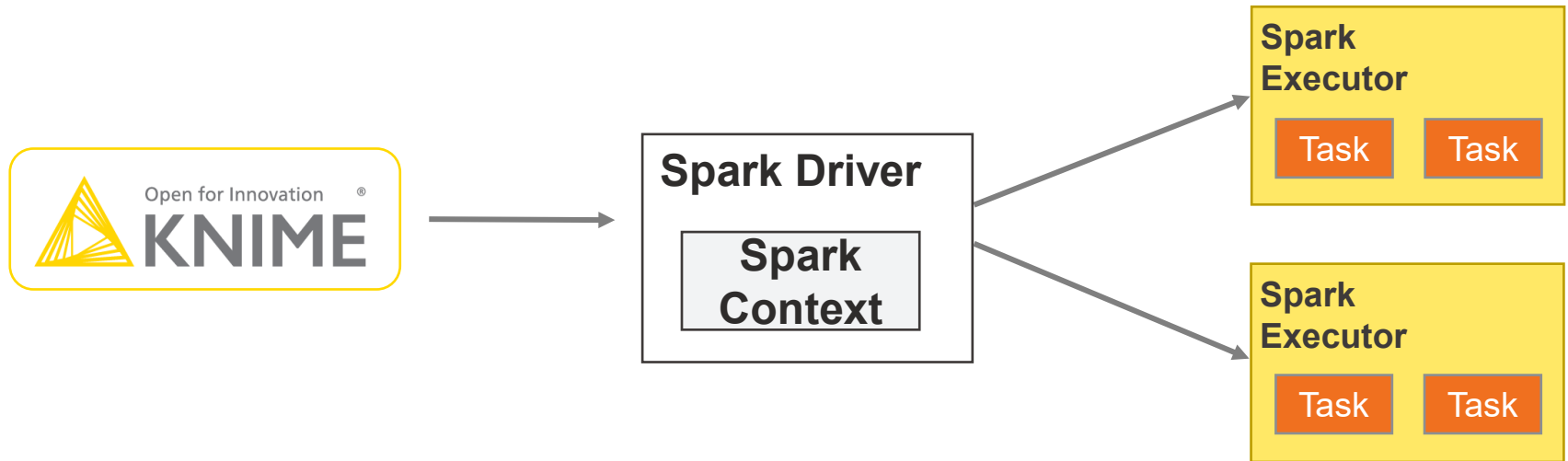
- Functions ("transformations") on DataFrames are not executed immediately
- Spark keeps record of the transformations for each DataFrame
- The actual execution is only triggered once the data is needed
- Offers the possibility to optimize the transformation steps





# Spark Context

- Main entry point for Spark functionality
- Represents connection to a Spark cluster
- Allocates resources on the cluster



# Agenda

---



Introduction to Hadoop and Spark



KNIME Big Data Connectors



KNIME Extension for Apache Spark



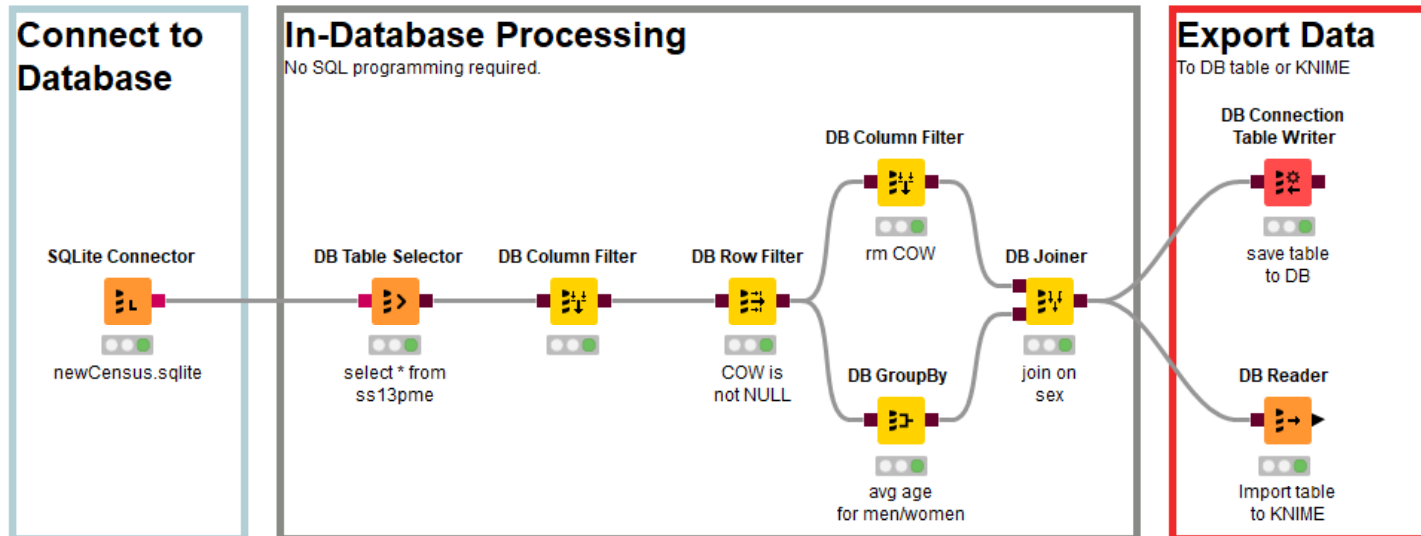
KNIME H2O Sparkling Water Integration



KNIME Workflow Executor for Apache Spark

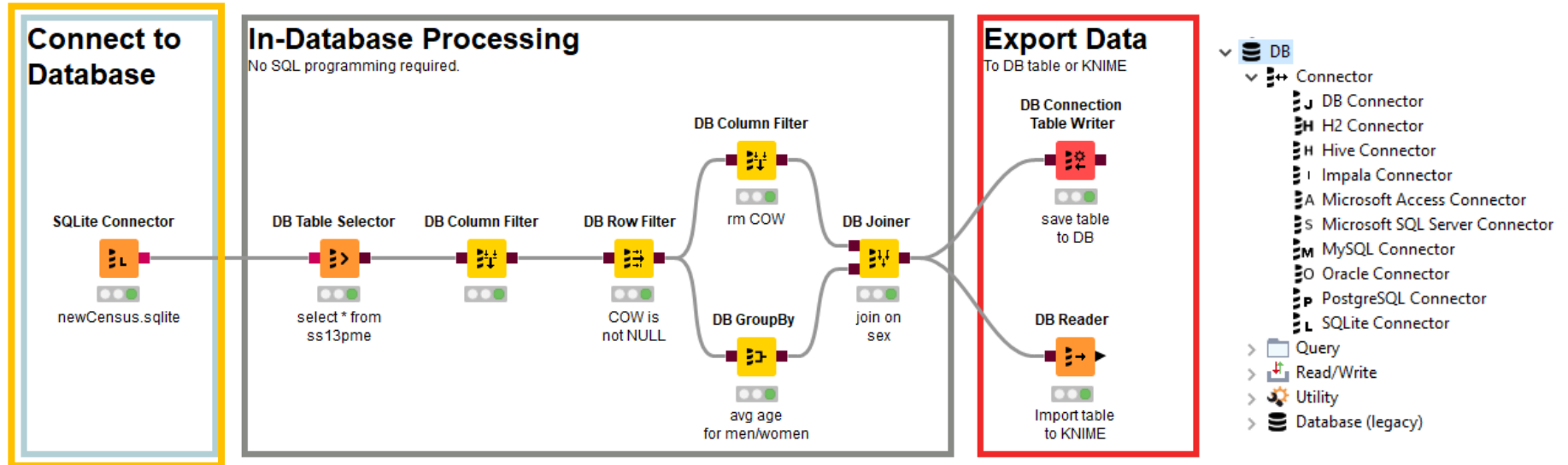
# Database Extension

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



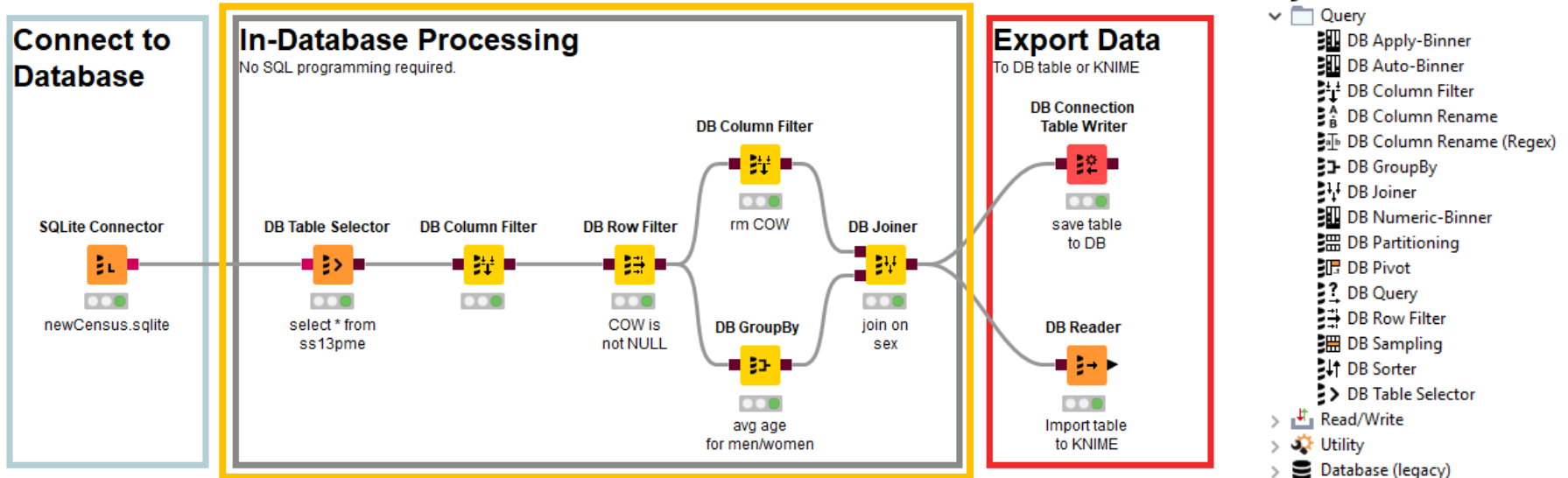
# Database Connectors

- Many dedicated DB Connector nodes available
- If connector node missing, use DB Connector node with JDBC driver



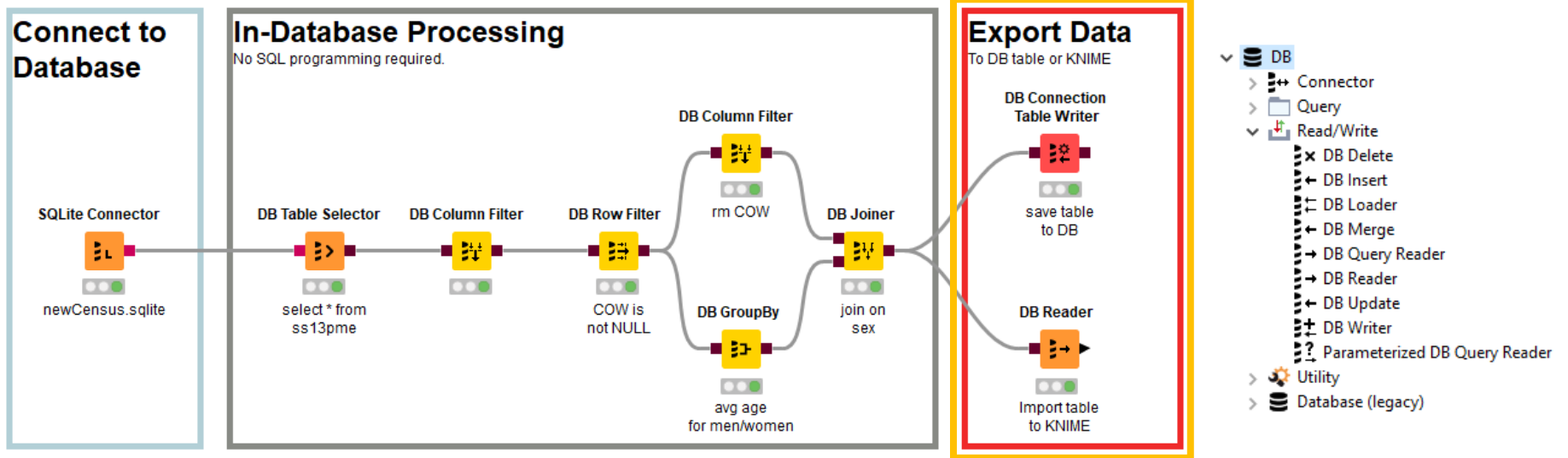
# In-Database Processing

- Database Manipulation nodes generate SQL query on top of the input SQL query (brown square port)
- SQL operations are **executed on the database!**



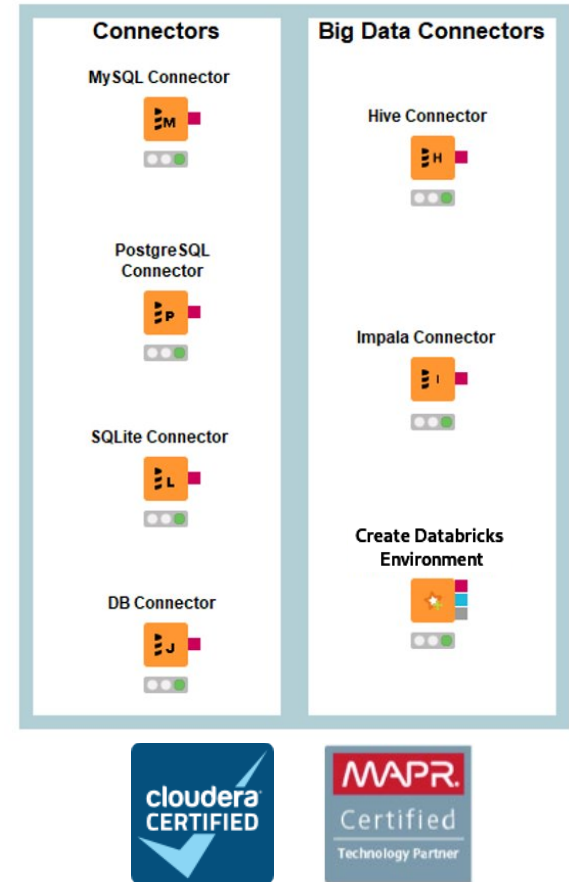
# Export Data

- Writing data back into database
- Exporting data into KNIME



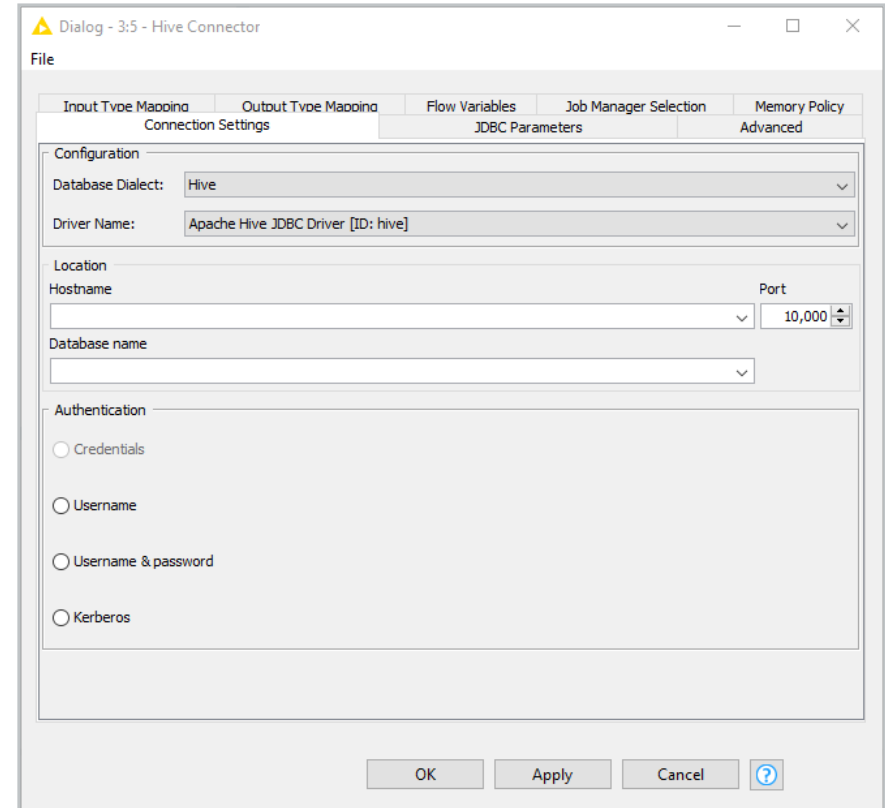
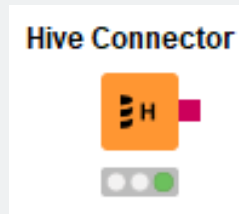
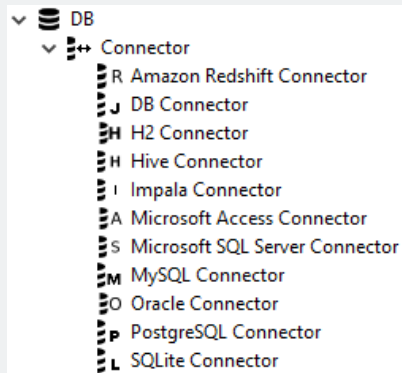
# KNIME Big Data Connectors

- Built upon Database extension
- Include drivers/libraries for HDFS, Hive, Impala and Databricks
- Preconfigured connectors
  - Hive
  - Impala
  - Databricks (Thriftserver)



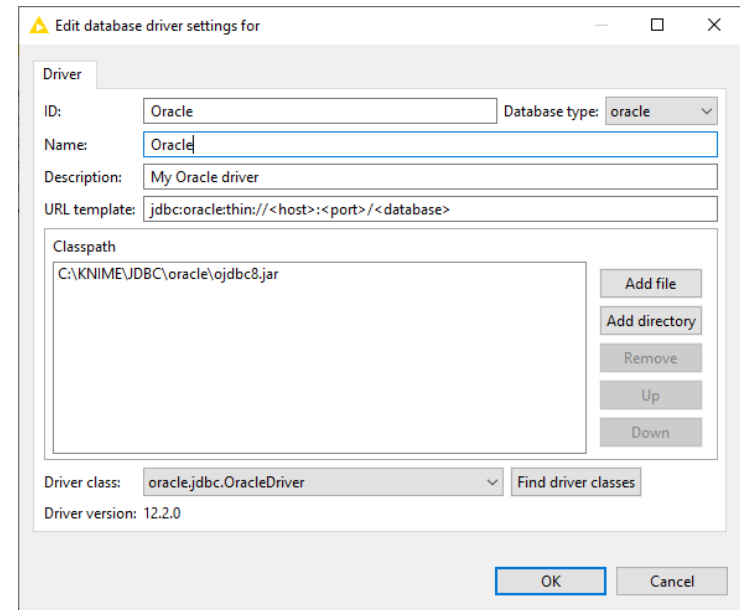
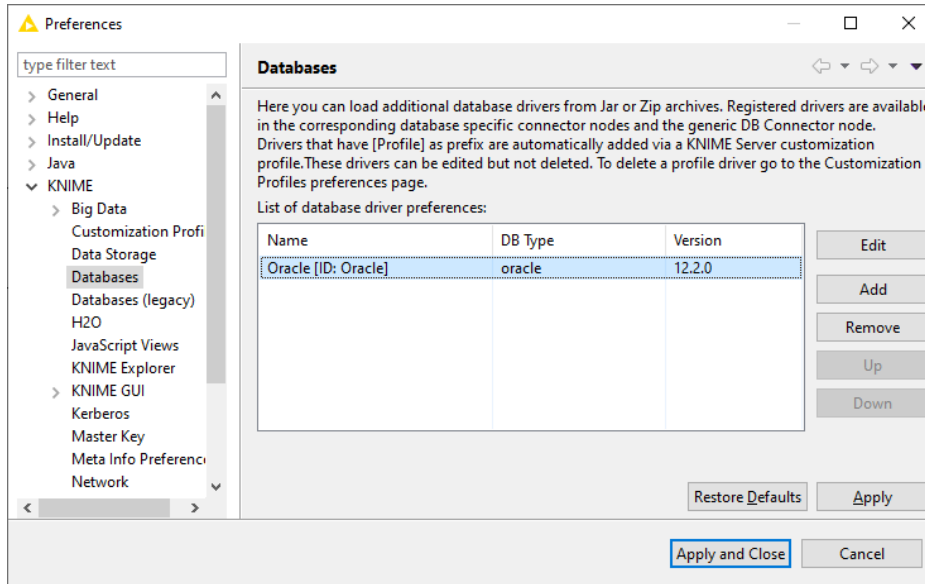
# Hive Connector

- Creates JDBC connection to Hive
- On unsecured clusters no password required





# Preferences: Registering proprietary JDBC drivers

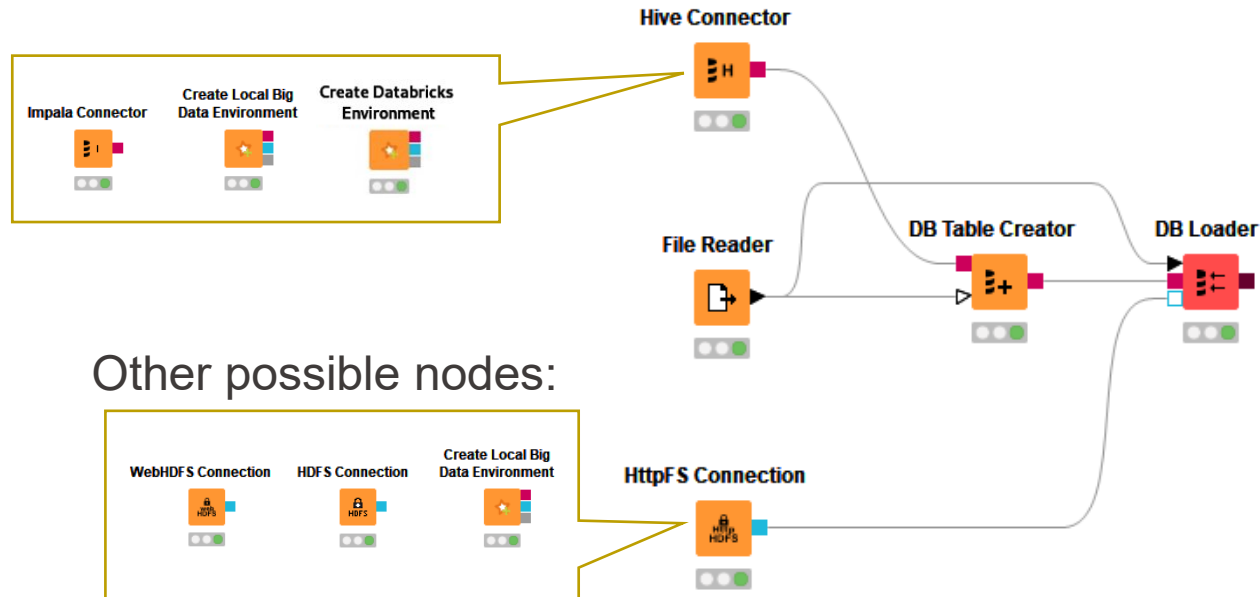


## Useful for:

- Cloudera Hive/Impala JDBC drivers
- Databricks JDBC Driver

# Loading Data into Hive/Impala

- Connectors are from KNIME Big Data Connectors Extension
- Use DB Table Creator and DB Loader from regular DB framework



# Demo: 01\_Flight\_Delay\_Statistics

Shortened URL to KNIME Hub folder: <https://tinyurl.com/yyjnenx8>

The screenshot displays the KNIME Hub interface. At the top left is the KNIME logo with the tagline "Open for Innovation". To its right is the word "Hub" and a search bar containing the text "Search workflows, nodes and more...". A user profile picture is visible in the top right corner. Below the navigation bar, the breadcrumb path is "KNIME Hub > tobias.koetter > Spaces > Public > Big\_Data\_Workshop\_2020", with a yellow arrow pointing to "Big\_Data\_Workshop\_2020". Below the breadcrumb, it says "Public space" and "Public". A second user profile picture is on the right. The main content area shows "Last update: 24 Jun 2020" and a heart icon with "0". Below this is a sub-navigation bar "Home > Big\_Data\_Workshop\_2020" with a refresh icon. A list of workflows follows:

←		
↻	<b>01_Flight_Delay_Statistics</b>	↻
↻	<b>02_Taxi_Demand_Prediction_Training_workflow</b>	↻

# Agenda

---



Introduction to Hadoop and Spark



KNIME Big Data Connectors



KNIME Extension for Apache Spark



KNIME H2O Sparkling Water Integration



KNIME Workflow Executor for Apache Spark

# KNIME Extension for Apache Spark

---



**Based on Spark MLlib**



**Scalable machine learning  
library**



**Supports algorithms for**

Classification

Regression

Clustering

Collaborative filtering

Dimensionality reduction



# Spark Integration in KNIME

- Apache Spark
  - IO
  - Column
  - Mining
  - Misc
  - Row
  - Statistics
    - Create Databricks Environment
    - Create Local Big Data Environment
    - Create Local Big Data Environment (Legacy)
    - Create Spark Context (Jobserver)
    - Create Spark Context (Livy)
    - Destroy Spark Context

- Apache Spark
  - IO
    - Database
      - DB to Spark
      - Database to Spark (legacy)
      - Hive to Spark
      - Hive to Spark (legacy)
      - Impala to Spark
      - Impala to Spark (legacy)
      - Spark to DB
      - Spark to Database (legacy)
      - Spark to Hive
      - Spark to Hive (legacy)
      - Spark to Impala
      - Spark to Impala (legacy)
    - Read
      - Avro to Spark
      - CSV to Spark
      - JSON to Spark
      - ORC to Spark
      - Parquet to Spark
      - Table to Spark
      - Text to Spark
    - Write
      - Spark to Avro
      - Spark to CSV
      - Spark to JSON
      - Spark to ORC
      - Spark to Parquet
      - Spark to Table
      - Spark to Text
    - Persist Spark DataFrame/RDD
    - Unpersist Spark DataFrame/RDD
  - 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
    - Transform
      - Spark Missing Value
      - Spark Missing Value (Apply)
      - Spark Column Filter

- Mining
  - Clustering
    - Spark Cluster Assigner
    - Spark k-Means
  - Dimensionality Reduction
    - Spark PCA
    - Spark SVD
  - Item Sets / Association Rules
    - Spark Association Rule (Apply)
    - Spark Association Rule Learner
    - Spark Frequent Item Sets
  - PMML
    - Spark Compiled Model Predictor
    - Spark MLLib to PMML
    - Spark PMML Model Predictor
  - Prediction
    - Spark Decision Tree Learner
    - Spark Decision Tree Learner (MLlib)
    - Spark Decision Tree Learner (Regression)
    - Spark Gradient Boosted Trees Learner
    - Spark Gradient Boosted Trees Learner (Regression)
    - Spark Gradient-Boosted Trees Learner (MLlib)
    - Spark Linear Regression Learner (MLlib)
    - Spark Linear SVM Learner (MLlib)
    - Spark Logistic Regression Learner (MLlib)
    - Spark Naive Bayes Learner (MLlib)
    - Spark Predictor (Classification)
    - Spark Predictor (MLlib)
    - Spark Predictor (Regression)
    - Spark Random Forest Learner
    - Spark Random Forest Learner (Regression)
    - Spark Random Forests Learner (MLlib)
  - Scoring
    - Spark Entropy Scorer
    - Spark Numeric Scorer
    - Spark Scorer
    - Spark Collaborative Filtering Learner (MLlib)

- Misc
  - Java Snippet
    - Spark DataFrame Java Snippet
    - Spark DataFrame Java Snippet (Sink)
    - Spark DataFrame Java Snippet (Source)
    - Spark RDD Java Snippet
    - Spark RDD Java Snippet (Sink)
    - Spark RDD Java Snippet (Source)
  - Management
    - List Spark DataFrames/RDDs
  - PySpark
    - PySpark Script (1 to 1)
    - PySpark Script (1 to 2)
    - PySpark Script (2 to 1)
    - PySpark Script (2 to 2)
    - PySpark Script Source
  - Spark Repartition
  - Spark SQL Query
  - Row
    - Spark Concatenate
    - Spark GroupBy
    - Spark Partitioning
    - Spark Pivot
    - Spark Row Filter
    - Spark Row Sampling
    - Spark Sorter
  - Statistics
    - Spark Correlation Filter
    - Spark Correlation Matrix
    - Spark Linear Correlation
    - Spark Statistics

# Spark Contexts: Creating

---

- **Create Local Big Data Environment**
  - Runs Spark locally on your machine (no cluster required)
  - Good for workflow prototyping
- **Create Spark Context (Livy)**
  - Requires a cluster that provides the Livy service
  - Good for production use
- **Create Databricks Environment**
  - Requires a Databricks cluster on AWS or Azure
  - Provides DB connection, DBFS and Spark

## Create Local Big Data Environment



Spark Context

## Create Spark Context (Livy)

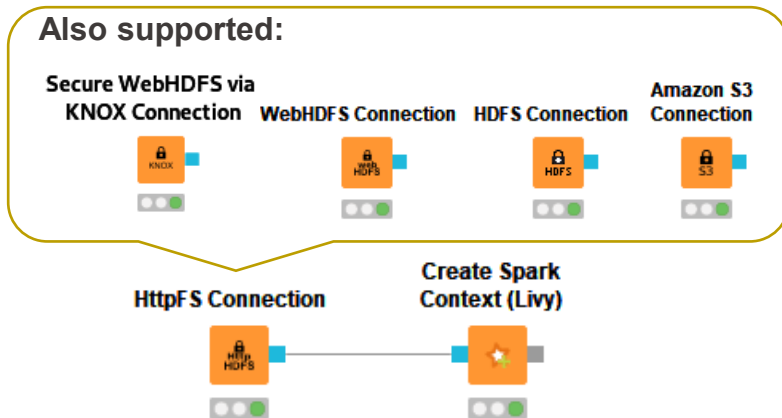
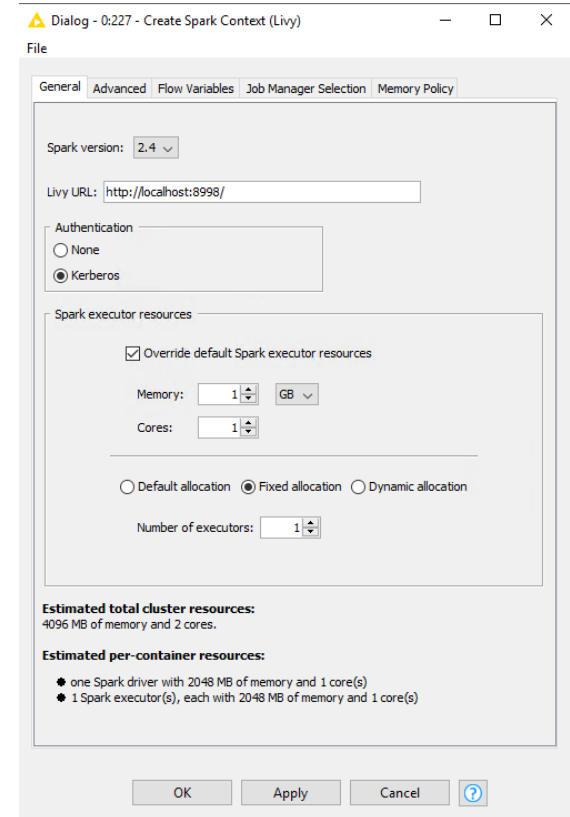


## Create Databricks Environment



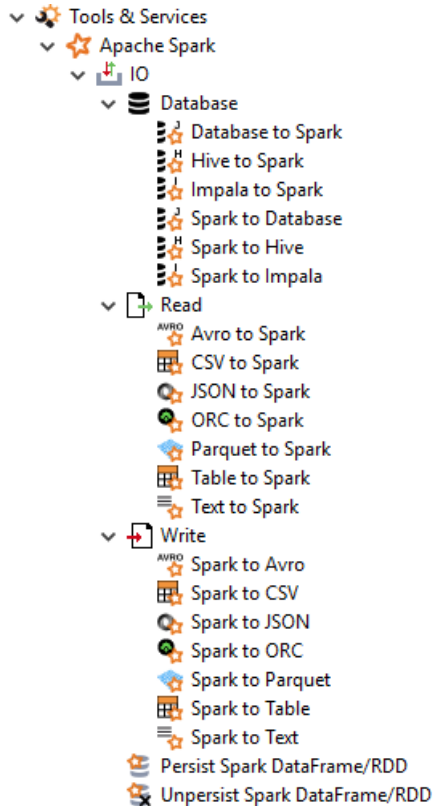
# Create Spark Context (Livy)

- Allows to use Spark nodes on clusters with Apache Livy
- Compatible with CDH, HDP, HDInsight and EMR

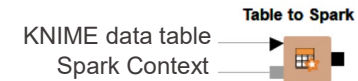




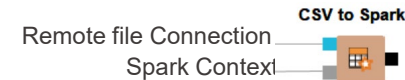
# Import Data to Spark



- From KNIME
- From CSV file in HDFS
- From Hive
- From other sources
- From Database



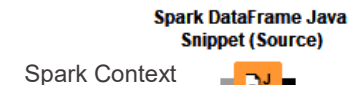
import a KNIME table to a Spark RDD



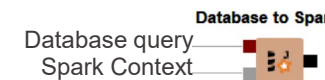
create a Spark RDD from a CSV file



convert a Hive query into a Spark RDD

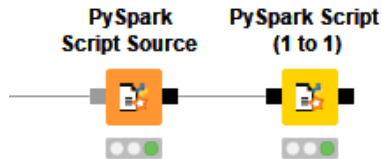
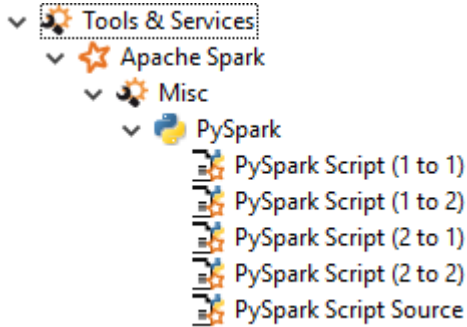


read file from HDFS



read database query into Spark RDD

# Modularize and Execute Your Own Spark Code: PySpark Script



The screenshot shows the 'Dialog - 3:9 - PySpark Script (1 to 1)' window. The 'Script' tab is active, displaying the following code:

```
1 # System imports
42 # Your custom imports:
43
44 # Flowvariables
45 # Your custom global variables:
46
47 # Initialization of Spark environment
55 # Custom pySpark code
56 # SparkSession can be used with variable spark
57 # The input dataframe(s): [dataFrame1]
58 # The output dataframe(s) must be: [resultDataFrame1]
59 resultDataFrame1 = dataFrame1.filter(dataFrame1.cow.isNotNull())
60
61
62
63
64 # End of user code
65 # Send data to jvm
67
```

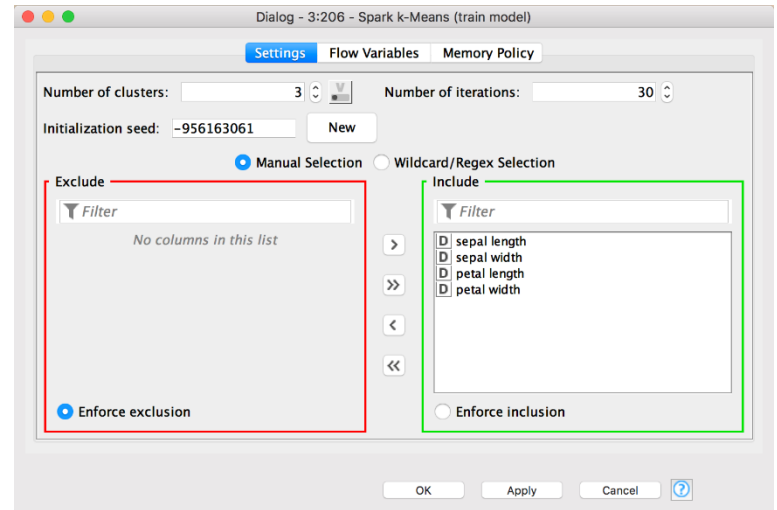
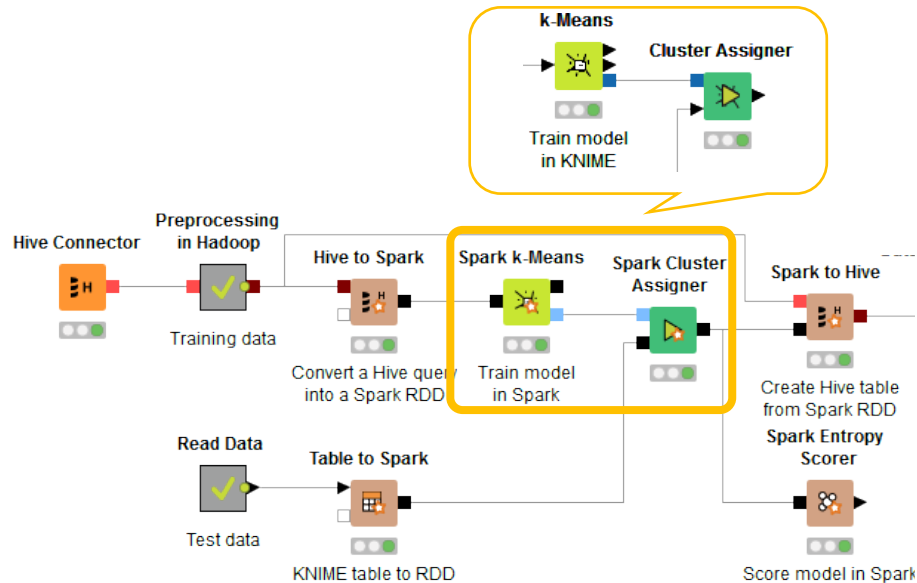
Below the code editor, there are buttons for 'Validate on Cluster' and a field for 'Number of rows to validate on' set to 50. The 'ResultDataFrame1(10 of 50 rows):' section shows a table of data:

serialno	rt/sporder	puma00	puma10	st	adjinc	pwgtp	agep	cit	citwp05	citwp12	cow	ddrs	dear	
12009000000425	P	1	900	-9	23	1085467	32	22	1	null	null	4	2	2
12009000001035	P	1	500	-9	23	1085467	23	57	1	null	null	6	2	2
12009000001051	P	1	900	-9	23	1085467	23	43	1	null	null	4	2	2
12009000001084	P	1	1000	-9	23	1085467	5	27	1	null	null	1	2	2
12009000001426	P	1	700	-9	23	1085467	17	47	1	null	null	1	2	2

At the bottom of the dialog, there are buttons for 'OK', 'Apply', 'Cancel', and a help icon.

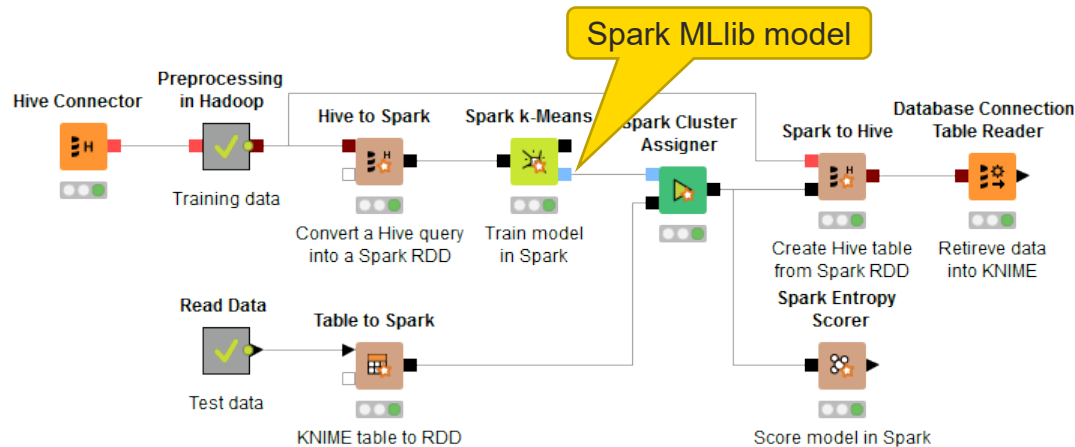
# MLlib Integration: Familiar Usage Model

- Usage model and dialogs like existing nodes
- No coding required
- Various algorithms for classification, regression and clustering supported



# MLlib Integration: Spark MLlib Model Port

- MLlib model ports for model transfer
- Model ports provide more information about the model itself



# Demo: 02\_Taxi\_Demand\_Prediction\_Training\_workflow

Shortened URL to KNIME Hub folder: <https://tinyurl.com/yyjnenx8>

The screenshot displays the KNIME Hub interface. At the top left is the KNIME logo with the tagline "Open for Innovation". To its right is the word "Hub" and a search bar containing the text "Search workflows, nodes and more...". On the far right of the top navigation bar is a user profile picture. Below the navigation bar is a breadcrumb trail: "KNIME Hub > tobias.koetter > Spaces > Public > Big\_Data\_Workshop\_2020". A large yellow arrow points to the "Big\_Data\_Workshop\_2020" link. Below the breadcrumb is a "Public space" icon and the word "Public". To the right of "Public" is another user profile picture. Below this is a section for the selected space, showing "Last update: 24 Jun 2020" and a heart icon with the number "0". Below this is a sub-breadcrumb trail: "Home > Big\_Data\_Workshop\_2020". A list of workflows is shown below, each with a left arrow icon, a workflow name, and a right arrow icon. The workflows listed are "01\_Flight\_Delay\_Statistics" and "02\_Taxi\_Demand\_Prediction\_Training\_workflow".

# Agenda

---



Introduction to Hadoop and Spark



KNIME Big Data Connectors



KNIME Extension for Apache Spark



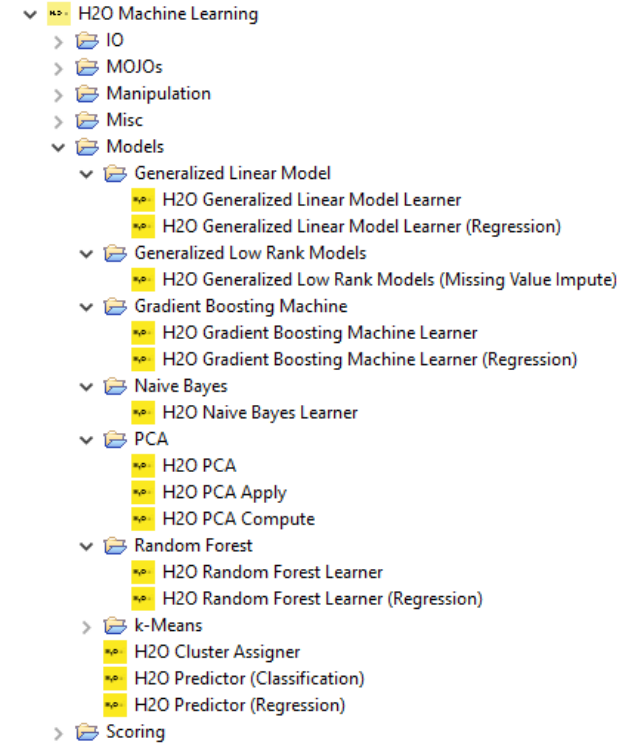
KNIME H2O Sparkling Water Integration



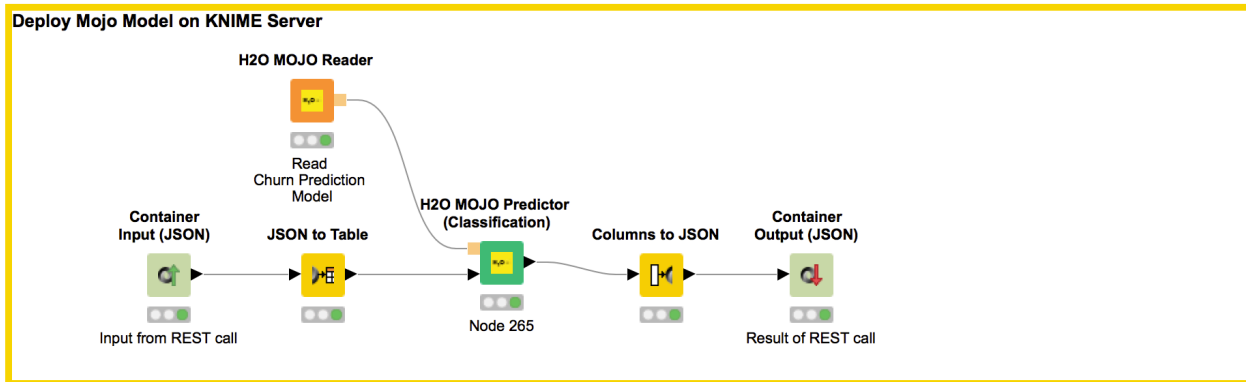
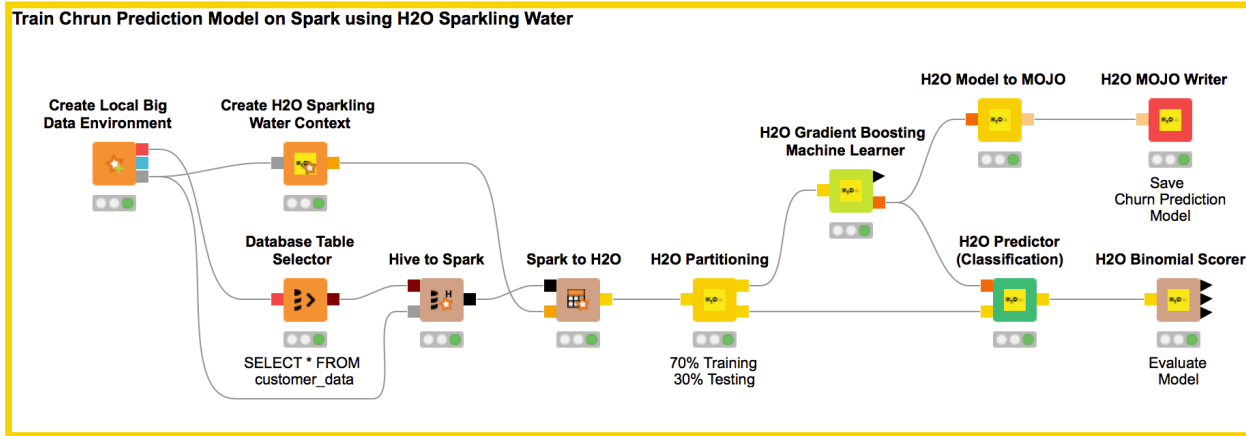
KNIME Workflow Executor for Apache Spark

# H2O Integration

- KNIME integrates the H2O machine learning library
- H2O: Open source, focus on scalability and performance
- Supports many different models
  - Generalized Linear Model
  - Gradient Boosting Machine
  - Random Forest
  - k-Means, PCA, Naive Bayes, etc.
- Includes support for MOJO model objects for deployment
- Sparkling water = H2O on Spark



# The H2O Sparkling Water Integration





# Agenda

---



Introduction to Hadoop and Spark



KNIME Big Data Connectors



KNIME Extension for Apache Spark

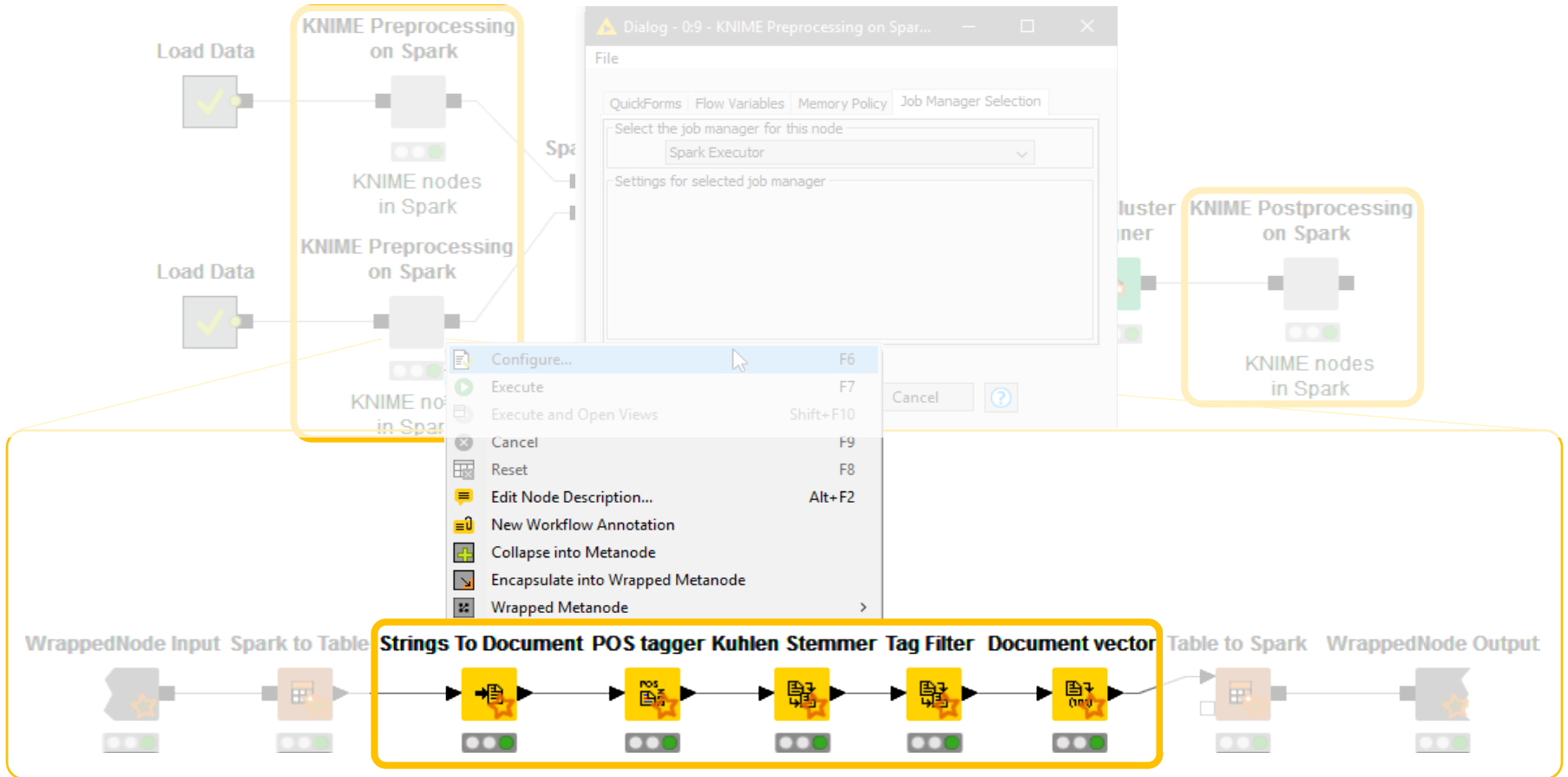


KNIME H2O Sparkling Water Integration



KNIME Workflow Executor for Apache Spark

# KNIME Workflow Executor for Apache Spark



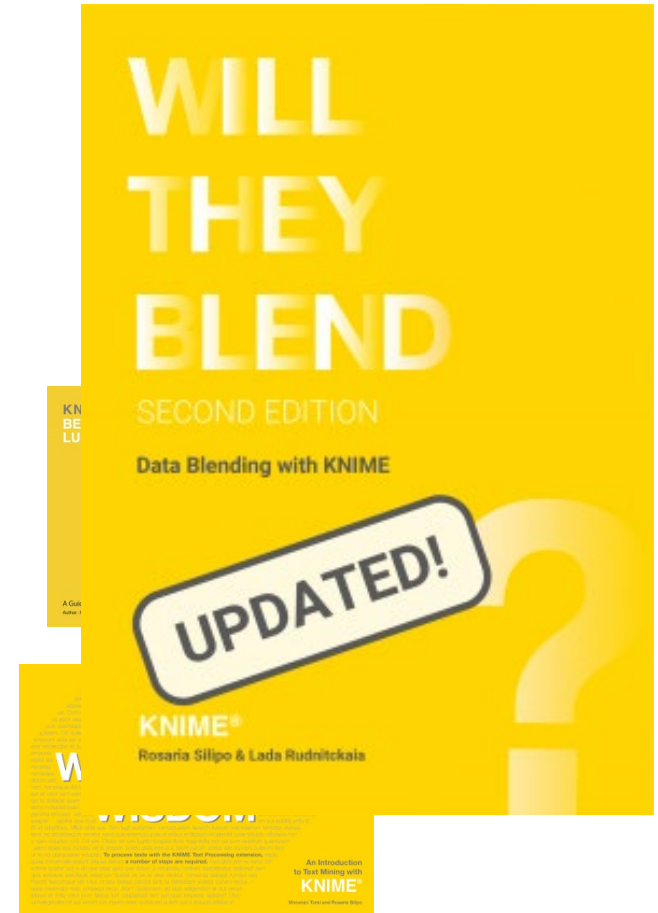
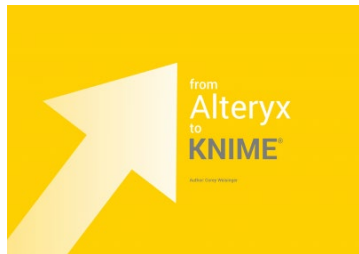
# Use Cases & Limitations

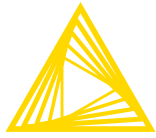
---

- Each workflow replica processes the rows of one partition!
- Good match for:
  - KNIME nodes that operate row-by-row
    - Many pre- and postprocessing nodes
    - Predictor nodes
    - Nodes that are streamable
  - Parallel execution of standard KNIME workflows on “small” data
    - Hyper-parameter optimization
- Bad match for: Any node that needs all rows, such as
  - GroupBy, Joiner, Pivoting, ...
  - Model learner nodes

# KNIME Books

- Course books downloadable from **KNIME Press**
- <https://www.knime.com/knimepress>
- Code: **FALL-SUMMIT-WORKSHOP**  
Valid for: All Books  
Expires: Jan 31, 2021





Open for Innovation

**KNIME**

**Thank you for joining!**

