KNIME Big Data Workshop

Tobias Kötter and Björn Lohrmann
KNIME
Variety, Volume, Velocity

• Variety:
  – integrating heterogeneous data...
  – .. and tools

• Volume:
  – from small files...
  – ...to distributed data repositories (Hadoop)
  – bring the computation to the data

• Velocity:
  – from distributing computationally heavy computations...
  – ...to real time scoring of millions of records/sec.
Variety
The KNIME Analytics Platform: Open for Every Data, Tool, and User

Data Scientist

Native Data Access, Analysis, Visualization, and Reporting

External Data Connectors

Distributed / Cloud Execution

External and Legacy Tools

Business Analyst

KNIME Analytics Platform

mongoDB

Facebook

Twitter

Hortonworks

Cloudera

Hive

IMPALA

TIBCO Spotfire

ACTUATE

BiRT

R

python

SSAS

SAS

AN IBM COMPANY

Open for Innovation

© 2016 KNIME.com AG. All Rights Reserved.
# Data Integration

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>5.4</td>
<td>3.9</td>
<td>1.7</td>
<td>0.4</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>4.6</td>
<td>3.4</td>
<td>1.4</td>
<td>0.3</td>
<td><em>I. setosa</em></td>
</tr>
<tr>
<td>5.0</td>
<td>3.4</td>
<td>1.5</td>
<td>0.2</td>
<td><em>I. setosa</em></td>
</tr>
</tbody>
</table>
Integrating R and Python
Other Programming/Scripting Integrations

Java Integrations
- Java Snippet
- Java Snippet Row Filter
- Java Snippet Row Splitter

JavaScript Integrations
- Generic JavaScript View
- JavaScript Table View
- JavaScript Scatter Plot
- JavaScript Lift Chart
- JavaScript Conditional Box Plot
- JavaScript ROC Curve
- JavaScript Line Plot
- JavaScript Box Plot
- JavaScript Bar Chart
- JavaScript Pie/Donut Chart

Misc Integrations
- External Tool
- External SSH Tool
- Generic Web Service Client
- SAS7DAT Reader
Volume
Database Extension

- Visually assemble complex SQL statements
- Connect to almost all JDBC-compliant databases
- Harness the power of your database within KNIME
In-Database Processing

- Operations are performed within the database
Tip

• SQL statements are logged in KNIME log file
Database Port Types

Database JDBC Connection Port (light red)
- Connection information

Database Connection Port (dark red)
- Connection information
- SQL statement

Database Connection Ports can be connected to Database JDBC Connection Ports but not vice versa
Database Connectors

• Nodes to connect to specific Databases
  – Bundling necessary JDBC drivers
  – Easy to use
  – 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 preferences page
Register JDBC Driver

Open KNIME and go to File -> Preferences

Increase connection timeout for long running database operations

Register new JDBC drivers
Query Nodes

- Filter rows and columns
- Join tables/queries
- Extract samples
- Bin numeric columns
- Sort your data
- Write your own query
- Aggregate your data
Returns number of rows per group
Database GroupBy – Pattern Based Aggregation

Tick this option if the search pattern is a regular expression otherwise it is treated as string with wildcards ('*' and '?')
Database GroupBy – Type Based Aggregation

Matches all columns
Matches all numeric columns
Database GroupBy – DB Specific Aggregation Methods

<table>
<thead>
<tr>
<th>Column</th>
<th>SQLite 7 Aggregation Functions</th>
<th>PostgreSQL 25 Aggregation Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>D sepal length</td>
<td>AVG</td>
<td>ARRAY AGG</td>
</tr>
<tr>
<td>D sepal width</td>
<td>AVG</td>
<td></td>
</tr>
<tr>
<td>D petal length</td>
<td>COUNT</td>
<td>AVG</td>
</tr>
<tr>
<td>D petal width</td>
<td>GROUP_CONCAT</td>
<td>CORR</td>
</tr>
<tr>
<td>S id</td>
<td>MAX</td>
<td>COUNT</td>
</tr>
<tr>
<td>D sepal width</td>
<td>MIN</td>
<td>COVAR_POP</td>
</tr>
<tr>
<td>D petal length</td>
<td>SUM</td>
<td>COVAR_SAMP</td>
</tr>
<tr>
<td>S id</td>
<td>TOTAL</td>
<td></td>
</tr>
<tr>
<td>S id</td>
<td>custom</td>
<td></td>
</tr>
<tr>
<td>D petal length</td>
<td>GROUP_CONCAT</td>
<td></td>
</tr>
</tbody>
</table>

© 2016 KNIME.com AG. All Rights Reserved.
Database GroupBy – Custom Aggregation Function

![Database GroupBy – Custom Aggregation Function](image)

The image shows the settings for database groupby with a focus on custom aggregation functions. The user interface includes options for setting up aggregation rules, with specific sections for groups, manual aggregation, pattern-based aggregation, and type-based aggregation. The screenshot demonstrates how to define a custom aggregation function using a dialogue box with fields for function definition, result column name, and optional second column. The interface also allows for advanced settings, including column naming and aggregation method selection.
Database Writing Nodes

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

– Increase batch size in database manipulation nodes
KNIME Big Data Connectors

• Package required drivers/libraries for HDFS, Hive, Impala access
• Preconfigured connectors
  – Hive
  – Cloudera Impala
  – Extends the open source database integration
Hive/Impala Loader

- Batch upload a KNIME data table to Hive/Impala
HDFS File Handling

• New nodes
  – HDFS Connection
  – HDFS File Permission

• Utilize the existing remote file handling nodes
  – Upload/download files
  – Create/list directories
  – Delete files
HDFS File Handling
KNIME Spark Executor

• Based on Spark MLlib
• Scalable machine learning library
• Runs on Hadoop
• Algorithms for
  – Classification (decision tree, naïve bayes, ...)
  – Regression (logistic regression, linear regression, ...)
  – Clustering (k-means)
  – Collaborative filtering (ALS)
  – Dimensionality reduction (SVD, PCA)
Familiar Usage Model

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

- MLlib model ports for model transfer
- Native MLlib model learning and prediction
- Spark nodes start and manage Spark jobs
  - Including Spark job cancelation
Data Stays Within Your Cluster

- Spark RDDs as input/output format
- Data stays within your cluster
- No unnecessary data movements
- Several input/output nodes e.g. Hive, hdfs files, ...
Machine Learning – Unsupervised Learning Example
Machine Learning – Supervised Learning Example
Mass Learning – Fast Event Prediction

- Convert supported MLlib models to PMML
Sophisticated Learning - Mass Prediction

- Supports KNIME models and pre-processing steps
Closing the Loop

Apply model on demand

Learn model at scale

Sophisticated model learning

Apply model at scale

PMML model

MLlib model

Open for Innovation

KNIME

Spark MLlib to PMML

Spark PMML Model Predictor
Mix and Match

- Combine with existing KNIME nodes such as loops
Modularize and Execute Your Own Spark Code
Lazy Evaluation in Spark

- Transformations are lazy
- Actions trigger evaluation
Spark Node Overview
KNIME Big Data Architecture

- **Scheduled execution and RESTful workflow submission**
  - KNIME Server with extensions:
    - KNIME Big Data Connectors
    - KNIME Big Data Executor for Spark
  - Workflow Upload via HTTP(S)
  - Build Spark workflows graphically

- **Submit Impala queries via JDBC**
  - Impala
  - Hadoop Cluster
  - Cloudera IMPALA
  - Hiveserver 2
  - *Software provided by KNIME, based on https://github.com/spark-jobserver/spark-jobserver*

- **Submit Hive queries via JDBC**
  - Hiveserver 2

- **Submit Spark jobs via HTTP(S)**
  - Spark Job Server
  - Spark

- **KNIME Analytics Platform with extensions:**
  - KNIME Big Data Connectors
  - KNIME Big Data Executor for Spark

- **Build Spark workflows graphically**

© 2016 KNIME.com AG. All Rights Reserved.
Velocity
Velocity

• Computationally Heavy Analytics:
  – Distributed Execution of one workflow branch
  – Parallel Execution of workflow branches
  – Parallel Execution of workflows
KNIME Cluster Executor: Distributed Data
KNIME Batch Executors on Hadoop (via Oozie)

- **KNIME Workflow Designer** designs and uploads workflow.
- **REST Client** submits workflow parameters via REST.
- **Oozie Server** executes workflows.
- **Hadoop Master Node** and **Hadoop Worker Node** manage the execution.
- **YARN** manages resources and containers.
- **HDFS** stores datasets.
- **YARN Container** executes KNIME workflows.
KNIME Executors on Hadoop (via YARN)

KNIME Workflow Users

Execute workflow

REST Client

KNIME Workflow Designer

YARN Container

Hadoop Worker Node

YARN

KNIME Server

 Execute workflow

Uploads

Submits

Starts

Executes

KNIME Executors on Hadoop (via YARN)
Velocity

• Computationally Heavy Analytics:
  – Distributed Execution of one workflow branch
  – Parallel Execution of workflow branches
  – Parallel Execution of workflows

• High Demand Scoring/Prediction:
  – High Performance Scoring using generic Workflows
  – High Performance Scoring of Predictive Models
High Performance Scoring via Workflows

• Record (or small batch) based processing
• Exposed as RESTful web service
High Performance Scoring using Models

- KNIME PMML Scoring via compiled PMML
- Deployed on KNIME Server
- Exposed as RESTful web service

- Partnership with Zementis
  - ADAPA Real Time Scoring
  - UPPI Big Data Scoring Engine
Velocity

• Computationally Heavy Analytics:
  – Distributed Execution of one workflow branch
  – Parallel Execution of workflow branches
  – Parallel Execution of workflows

• High Demand Scoring/Prediction:
  – High Performance Scoring using generic Workflows
  – High Performance Scoring of Predictive Models

• Continuous Data Streams
  – Streaming in KNIME
  – Spark Streaming
Streaming in KNIME
Spark Streaming
Big Data, IoT, and the three V

• Variety:
  – KNIME inherently well-suited: open platform
  – broad data source/type support
  – extensive tool integration

• Volume:
  – Bring the computation to the data
  – Big Data Extensions cover ETL and model learning

• Velocity:
  – Distributed Execution of heavy workflows
  – High Performance Scoring of predictive models
  – Streaming execution
Demo
Virtual Machines

- Hortonworks:  
- Cloudera:  
- KNIME VM based on Hortonworks sandbox with preinstalled Spark Job Server is coming soon
Resources

– SQL Syntax and Examples (www.w3schools.com)
– Apache Spark MLlib (http://spark.apache.org/mllib/)
– The KNIME Website (www.knime.org)
  • Database Documentation (https://tech.knime.org/database-documentation)
  • Big Data Extensions (https://www.knime.org/knime-big-data-extensions)
  • Forum (tech.knime.org/forum)
  • LEARNING HUB under RESOURCES (www.knime.org/learning-hub)
  • Blog for news, tips and tricks (www.knime.org/blog)

– KNIME TV channel on YouTube
– KNIME on Twitter @KNIME
The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME.com AG under license from KNIME GmbH, and are registered in the United States.