What's New and Cooking in the KNIME Big Data Labs

Tobias Kötter and Björn Lohrmann

KNIME
Recap

Database Integration and KNIME Big Data Connectors
Database Integration - Recap

- Visually assemble complex SQL statements
- Connect to almost all JDBC-compliant databases
- Preconfigured nodes to connect to various databases
- Harness the power of your database within KNIME
KNIME Big Data Connectors - Recap

• Package required drivers/libraries for HDFS, Hive, Impala access
• Performs operations on Hadoop
• Extends the open source database integration
• Preconfigured connectors
  – Hive
  – Cloudera Impala
What’s New

Database Integration and
KNIME Big Data Connectors
New Database Nodes

- Python Script (DB)/(Hive)
- Database Nummeric-/Auto-Binner and Apply-Binner
- Database Sampling with support for stratified sampling
- Database Pivot
What’s New

KNIME Spark Executor
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
- Spark nodes start and manage Spark jobs
- No coding required
In-Hadoop Processing

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

Mix and Match

- Read Data
- Normalizer (PMML)
- R Learner
- R To PMML
- Spark Transformations Applier
- Spark PMML Model Predictor
- Spark Scorer
- HeatMap

- Training Data
- Preprocessing in Hive
- Hive to Spark
- Apply PMML transformations in Spark
- Mass prediction in Spark
- Score model in Spark

- Hive Connector
- Test Data
- Hive query to Spark RDD
Let KNIME Control Your Spark Jobs
47 Spark Nodes and Counting
More than 170 Mio rows with energy usage data from smart meters

Uses KNIME Analytics Platform, Big Data Connectors and Spark Executor to forecast energy consumption
What's Cooking

Database Integration and KNIME Big Data Connectors
Database Integration

• Improved ...
  – connection handling
  – schema support
  – database driver handling
  – type handling with support for arrays

• New node:
  – Table Creator with support for unique/primary keys
KNIME Big Data Connectors

• Improved secured cluster support
  – Full scale Kerberos support
  – Apache KNOX integration

• New nodes:
  – Amazon S3
  – Amazon Redshift
  – Phoenix
  – HBase
  – Drill
What's Cooking

KNIME Spark Executor
KNIME Spark Executor Development

- Extended Spark version support
- Public API to write your own Spark nodes
- Enhanced support for multiple Spark contexts
- Support for secured clusters

- New nodes:
  - GroupBy
  - Remote File Reader/Writer
  - Database to Spark/Spark to Database
Virtual Data Warehouse

Load from different sources
- Hive Connector
- MySQL Connector
- Database Connector
  - any JDBC compliant DB
    - Read Data
  - Local Pre-processing

Preprocessing in Hive
In-database processing

On the fly combining, preprocessing and analysis in Spark
- Hive to Spark
- Spark Category
  - To Number
- Spark Joiner
- Spark Normalizer
- Spark Column Rename
- Spark Concatenate
- Spark Java Snippet
- Spark to Database
- Spark to Hive
- Spark to Table

Deploy to different targets
- Hive Connector
- Spark to Hive
- Spark to Table
- Database Connector
  - any JDBC compliant DB
    - Read Data
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.