Harnessing Big Data with KNIME

Tobias Kötter
KNIME.com
Agenda

• The three V’s of Big Data
• Big Data Extension and Databases Nodes
• Demo
Variety:
• integrating heterogeneous data (and tools)

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

Velocity:
• from distributing computationally heavy computations...
• ...to real time scoring of millions of records/sec.
Variety
Variety

• Data Integration
  – Small (Ascii)
  – Proprietary (XLS, SAS...)
  – Medium (Databases)
  – Large (Hive, Impala, ParStream, HP Vertica...)
  – Diverse (Numbers, Texts, Images, Networks, Sequences...)

• Tool Integration
  – Native
  – Legacy, Inhouse
  – R, Python, Matlab, ...
Volume
Every Minute...

- Twitter users send out 277,000 tweets
- Google processes more than 2 million search queries
- Facebook processes almost 350 GB of data
- 72 hours of new video are uploaded to YouTube
- Individuals and organizations launch 571 new websites
- Walmart processes almost 17,000 transactions
- More than 100 million new emails are generated
- Sprint processes more than 250,000 phone calls
THE INTERNET OF THINGS
AN EXPLOSION OF CONNECTED POSSIBILITY

1982
1,000,000
- Apple's First Macintosh

1988
0.5 BILLION
- Dial-up Modems Popular

2003
IoT INCEPTION
- Nest and Black & White

2010
22.9 BILLION
- Sensor data in a smart home

2015
18.2 BILLION
- Wearables taking off

2016
34.8 BILLION
- Internet of Everything

2017
28.4 BILLION
- Internet of Things

2018
11.2 BILLION
- AI and IoT Convergence

2019
42.1 BILLION
- Internet of Everything

2020
50.1 BILLION
- Internet of Things

2025
100 BILLION
- Internet of Everything

- IoT is predicted to reach 50 billion devices by 2020

- IoT is an explosion of connected possibility

- IoT is predicted to reach 100 billion devices by 2025

- IoT is predicted to reach 1 Trillion devices by 2040
Big Data Support

• KNIME Database Nodes
  – in database processing
  – preconfigured connectors

• KNIME Big Data Extension
  – package required drivers/libraries for specific HDFS, Hive, Impala access

• Spark MLlib integration (coming soon)
Velocity
Velocity

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

• Hosted Analytics/Prediction
  – Web service Deployment of Workflows

• High Demand Scoring/Prediction:
  – Continuous Execution of Workflow parts
  – High Performance Scoring using generic Workflows
  – High Performance Scoring of Predictive Models
KNIME Cluster Execution: Distributed Data
KNIME Cluster Execution: Distributed Analytics
Deployed Workflows

Application Access
- Custom API
- WSDL/SOAP based
Continuous Scoring using Workflows

- Exposes workflow fragment as RESTful web service
- Deployed on KNIME Server (v4.0 – 1H2015)
High Performance Scoring via Workflows

- Streaming Executor
- Deployed via KNIME Server (v4.1 – 2H2015/2016)
- Record (or small batch) based processing
- Exposed as RESTful web service
High Performance Scoring using Models

- Deployed on KNIME Server (v4.0 – 1H2015)
- KNIME PMML Scoring via compiled PMML
- Exposed as RESTful web service

- Partnership with Zementis
  - ADAPA Real Time Scoring
  - UPPI Big Data Scoring Engine
Big Data, IoT, and the three V

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

Volume:
- Big Data Extensions cover Hadoop based data integration and aggregation
- *Big Data Executors will address model building and streaming execution*

Velocity:
- Distributed Execution of heavy workflows to...
- *High Performance Scoring of predictive models.*
Big Data Extension and Database Nodes
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 JDBC Connection Port View

- Database Driver: org.postgresql.Driver
- Database URL: jdbc:postgresql://192.168.56.111:5432/
- User Name: knime
- Database Type: postgresql
Database Connection Port View

Select the table

- Configure
- Execute
- Execute and Open Views
- Cancel
- Reset
- Edit Node Description
- New Workflow Annotation
- Collapse into Meta Node
- Collapse into Sub Node
- Show Flow Variable Ports
- Cut
- Copy
- Paste
- Undo
- Redo
- Delete

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
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 Big Data 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
Reader/Writer

- Table selection
- Load data into KNIME
- Create table as select
- Insert/append data
- Delete rows from table
- Update values in table
Hive/Impala Loader

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

Partitioning influences performance
Manipulation

- Filter rows and columns
- Join tables/queries
- Sort your data
- Write your own query
- Aggregate your data
Database GroupBy – DB Specific Aggregation Methods

SQLite 7 aggregation functions

PostgreSQL 25 aggregation functions
### Database GroupBy – Aggregation Method Description

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGR_INTERCEPT</td>
<td>The function <code>regr_intercept(Y, X)</code> returns the y-intercept of the least-squares-fit linear equation determined by the (X, Y) pairs.</td>
</tr>
<tr>
<td>REGR_R2</td>
<td>The function <code>regr_r2(Y, X)</code> returns the square of the correlation coefficient.</td>
</tr>
<tr>
<td>REGR_SLOPE</td>
<td>The function <code>regr_slope(Y, X)</code> returns the slope of the least-squares-fit linear equation determined by the (X, Y) pairs.</td>
</tr>
<tr>
<td>REGR_SXX</td>
<td>The function <code>regr_sxx(Y, X)</code> returns the 'sum of squares' of the independent variable (sum(X^2) - sum(X)(^2)/N).</td>
</tr>
<tr>
<td>REGR_SXY</td>
<td>The function <code>regr_sxy(Y, X)</code> returns the 'sum of products' of independent times dependent variable (sum(X*Y) - sum(X) * sum(Y)/N).</td>
</tr>
<tr>
<td>REGR_SYX</td>
<td>The function <code>regr_syx(Y, X)</code> returns the 'sum of squares' of the dependent variable (sum(Y^2) - sum(Y)(^2)/N).</td>
</tr>
<tr>
<td>STDDEV_POP</td>
<td>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.</td>
</tr>
<tr>
<td>STDDEV_SAMP</td>
<td>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).</td>
</tr>
</tbody>
</table>
Database GroupBy – Manual Aggregation

Returns number of rows per group
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 cells
- Matches all numeric cells
Database GroupBy – Custom Aggregation Function

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

The image shows a dialogue box for setting up a custom aggregation function in a database. The function is defined as `FUNCTION(#COLUMN_NAME#)` and the result column name is set to `CUSTOM`. There is an option for an optional second column set to `<none>`. Advanced settings include column naming and aggregation method. The aggregation method includes an option to add `COUNT(*)` as a column name.
Utility

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

Executes several queries separated by `;` and new line
In-Database Processing

Loads your pre-processed data into KNIME
HDFS File Handling

- KNIME & Extensions -> KNIME File Handling Nodes
- HDFS Connection and HDFS File Permission nodes part of the commercial Big Data Extension
HDFS File Handling
Virtual Machines

• Hortonworks:  
  http://hortonworks.com/products/hortonworks-sandbox/

• Cloudera:  

• Virtual Box  
  https://www.virtualbox.org/

• VMWare Player  
  https://www.virtualbox.org/
Demo
Resources

- **KNIME** ([www.knime.org](http://www.knime.org))
- **BLOG** for news, tips and tricks ([www.knime.org/blog](http://www.knime.org/blog))
- **FORUM** for questions and answers ([tech.knime.org/forum](http://tech.knime.org/forum))
- **EXAMPLE SERVER** for example workflows
- **LEARNING HUB** ([www.knime.org/learning-hub](http://www.knime.org/learning-hub))

- **KNIME TV** channel on [YouTube](https://www.youtube.com/c/Knime)
- **KNIME** on [Twitter](https://twitter.com/Knime) @KNIME
- **KNIME** on [Facebook](https://www.facebook.com/KNIMEanalytics)