Integrating Big Data is as easy as 1,2,3 ... 4!

Tobias Kötter
KNIME.com
Variety, Volume, Velocity
Variety, Volume, Velocity

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.
Every Minute...

Twitter users send out 277,000 tweets
Facebook processes almost 350 GB of data
Every Minute...
Google processes more than 2 million search queries
72 hours of new video are uploaded to YouTube
Individuals and organizations launch 571 new websites
Sprint processes more than 250,000 phone calls
Walmart processes almost 17,000 transactions
More than 100 million new emails are generated

The Challenge
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

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

Workflow 3
Workflow 1: PrepareData

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

- Read all Data
  - Read 6 files for a total of 176 Mio Rows
- String to datetime
  - convert proprietary date format into datetime values
- 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
- Joiner
  - % values
  - intra-day and intra-week kW % usage by meter ID
  - Node 50
- 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
- Write to server
- Write to CSV
- Node 99
- Node 100

≈ 2 days
Big Data
Big Data Support

• KNIME Big Data Access Nodes
  – in database processing
  – preconfigured connectors

• Big Data Platforms
  – HDFS, Hive, Impala, HP Vertica, Hortonworks, ParStream, any big data platform really!

• Spark MLlib integration (coming soon)
• Streaming Executor (coming soon)
Virtual Machines

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

• Cloudera:

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

• VMWare Player
  http://www.vmware.com/
Accessing Big Data: Database Connector

Generic 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 retrieval operations
Accessing Big Data: Dedicated Connectors

Dedicated pre-configured connectors
- Bundling necessary JDBC drivers
- Easy to use
- DB specific behavior/capability

Some dedicated connectors are part of the KNIME Analytics Platform, some belong to the commercial KNIME Big Data Extension

works for most Hadoop HIVE installations, including Hortonworks

free
Accessing Big Data: Dedicated Connectors

Dedicated Connector

[Image of a dedicated connector setup dialog]

License: This material is protected by copyright. No reuse without license.
Data Table Selection

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

Database Table Selector

SQL Statement

SELECT * FROM energyDate
In-Database Processing
Manipulation

• Filter rows and columns
• Join tables/queries
• Sort your data
• Write your own query
• Aggregate* your data

* Database GroupBy node exposes DB specific aggregation methods
Adding SQL Queries for average Measures

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

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 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
This metanode calculates the energy usage time series aggregating the raw data at the yearly, monthly, weekly, daily, and hourly level.
Average Monthly Values
Import Data from Database into KNIME

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

SQL queries to calculate Energy Usage Measures
- daily, hourly, weekly, monthly, yearly, by intra-day segments
- 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

Run the SQL queries and retrieve final data
New Big Data Platform?

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.

No problem! Just change the connector node!
Other Useful Database Nodes

- **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.
KNIME Big Data Extension
KNIME Big Data Extension

• KNIME Big Data Access Nodes
  – preconfigured connectors
  – HDFS File Handling
  – Hive/Impala Loader

• Big Data Platforms
  – HDFS, Hive, Impala, HP Vertica, Hortonworks, ParStream, SAP Hana (to be), Teradata (to be), ...

• Spark MLlib integration (coming soon)
• Streaming Executor (coming soon)
HDFS File Handling

• KNIME & Extensions -> KNIME File Handling Nodes
• HDFS Connection and HDFS File Permission nodes
Hive/Impala Loader

- Upload a KNIME data table to Hive/Impala
KNIME Big Data Extension: Download and Install

KNIME.com Extension Store

License Required!

Installation Instructions
http://tech.knime.org/installation-instructions

Product Description
http://www.knime.org/knime-big-data-extension
License on KNIME Store

http://tech.knime.org/knime-store

These are the products you can purchase in this online store:

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNIME Big Data Extension</td>
<td>EUR 500.00 + tax per user per year</td>
</tr>
<tr>
<td>KNIME Personal Productivity</td>
<td>EUR 300.00 + tax per user per year</td>
</tr>
</tbody>
</table>

30-day trial license available with special Promotion Code education@knime.com
Thank You!

- education@knime.com

- We are hiring
  - Java Hadoop/BigData developers!
  - Senior Software Engineer - Web Development

- Whitepaper “KNIME opens the Doors to Big Data”

- Blog Post “Integrating Big data is as Easy as 1,2,3,4”
  http://www.knime.org/blog/integrating-big-data-is-as-easy-as-1-2-3-4