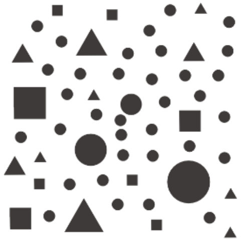
Open for Innovation
KNIME

Re-engineering IoT Legacy Analytics Solutions with Big Data

Rosaria Silipo & Bernd Wiswedel
KNIME.com

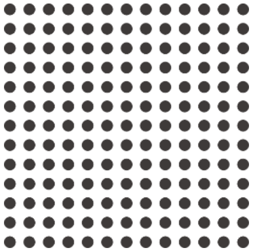
Rosaria.Silipo@knime.com

Variety, Volume, Velocity



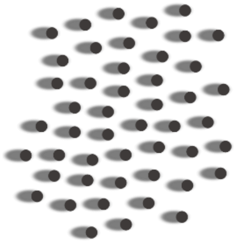
Variety:

- integrating heterogeneous data (and tools)



Volume:

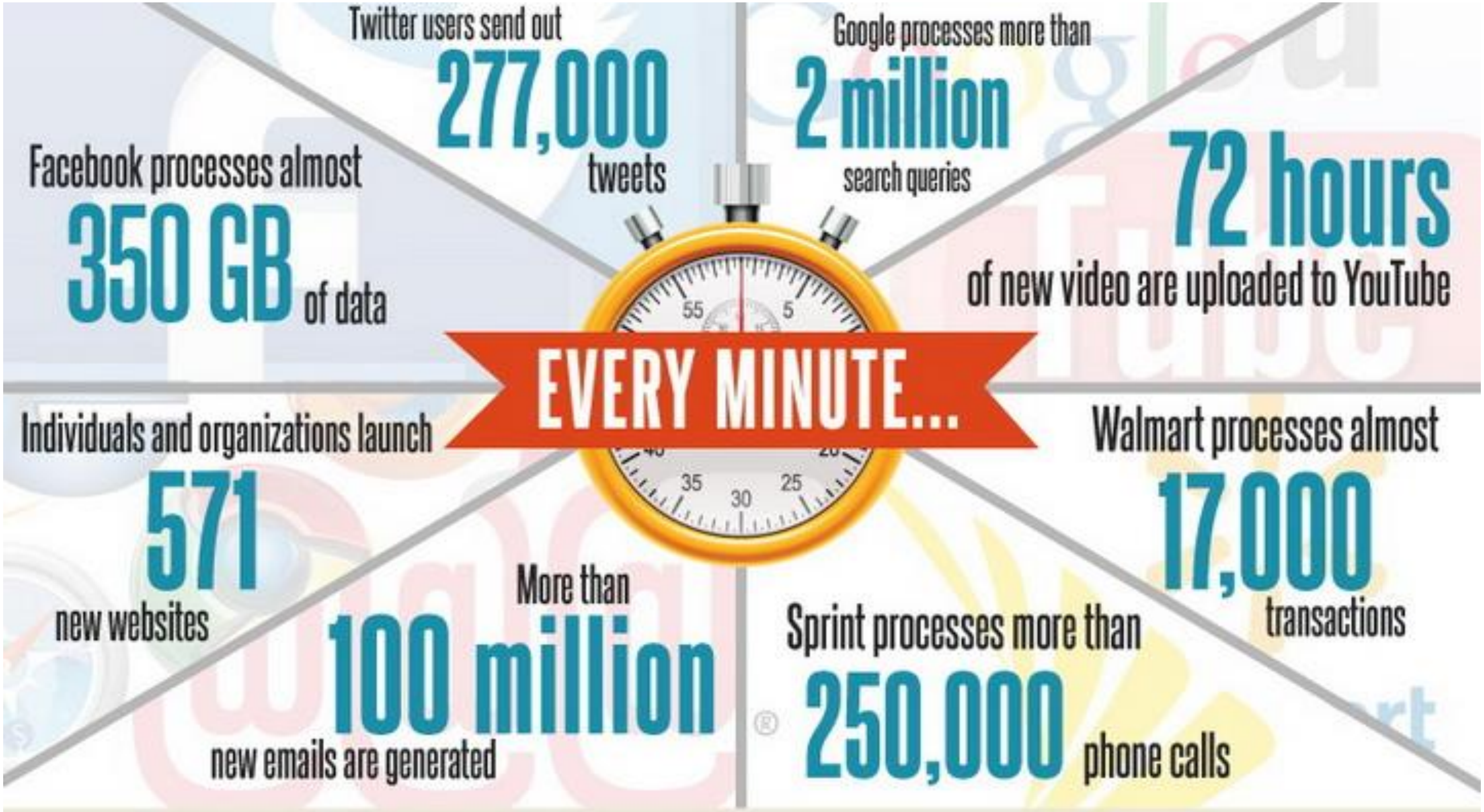
- from small files...
- ...to distributed data repositories (Hadoop)
- bring the tools to the data



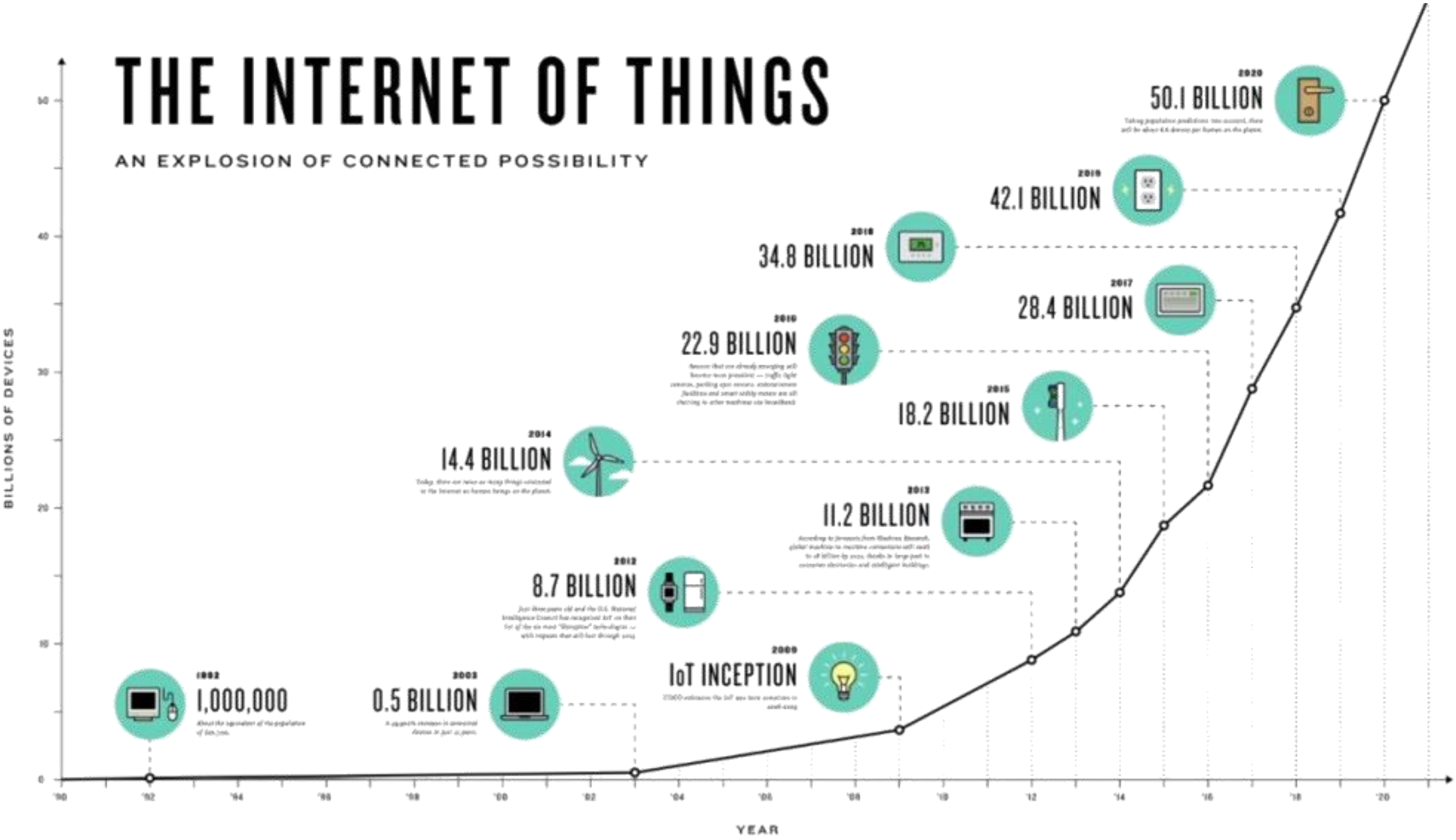
Velocity:

- from distributing computationally heavy computations...
- ...to real time scoring of millions of records/sec.

Every Minute…



IoT


THE INTERNET OF THINGS
AN EXPLOSION OF CONNECTED POSSIBILITY
BILLIONS OF DEVICES
0
10
20
30
40
50
YEAR
'90
'92
'94
'96
'98
'00
'02
'04
'06
'08
'10
'12
'14
'16
'18
'20
1992
1,000,000
About the equivalent of the population of San Jose.
2003
0.5 BILLION
A 499x increase in connected devices in just 11 years.
2009
IoT INCEPTION
[illegible]
2012
8.7 BILLION
Just three years old and the U.S. National Intelligence Council has recognized IoT on their list of the six most "disruptive" technologies — with impacts that will last through 2025.
2013
11.2 BILLION
According to forecasts from Machina Research, global machine-to-machine connections will swell to 18 billion by 2022, thanks in large part to consumer electronics and intelligent buildings.
2014
14.4 BILLION
Today, there are twice as many things connected to the Internet as human beings on the planet.
2015
18.2 BILLION
2016
22.9 BILLION
Sensors that are already emerging will become more prevalent — traffic light cameras, parking spot sensors, entertainment facilities and smart utility meters are all chatting to other machines via broadband.
2017
28.4 BILLION
2018
34.8 BILLION
2019
42.1 BILLION
2020
50.1 BILLION
Taking population predictions into account, there will be about 6.6 devices per human on the planet.

The IoT Legacy

Energy Usage Prediction from Smart Meters Data

- Read Smart Meter Energy Data (**176 millions rows**)
- Clean Up and Aggregate total Energy Usage by hour, week, day, month, year
- Calculate Behavioral Measures for each Smart Meter

Workflow 1

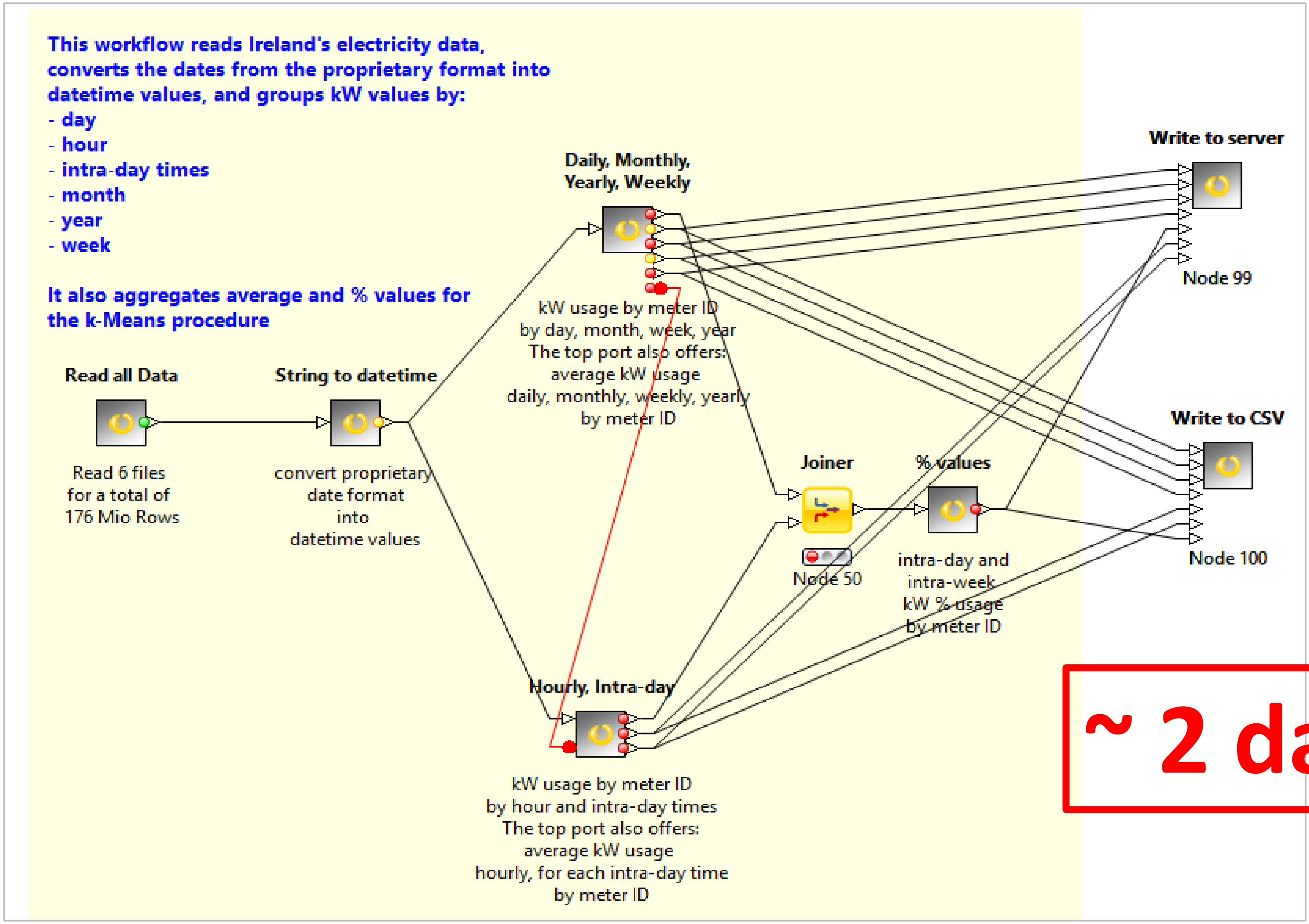
- Cluster Smart Meters with Similar Behavior (k-Means)

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

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

% values: intra-day and intra-week kW % usage by meter ID

~ 2 days

Big Data Options

Big Data Support

- KNIME Big Data Access Nodes
  - preconfigured connectors
  - in database processing
- Big Data Platforms
  - HDFS, Hive, Impala, HP Vertica, Hortonworks, ParStream, Actian, MapR, any big data platform really!
- Spark MLlib integration (coming soon)
- Streaming Executor (coming soon)

cloudera CERTIFIED

Hortonworks CERTIFIED TECHNOLOGY PARTNER

Hadoop Sandboxes

- Hortonworks:

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

- Cloudera:

  http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms.html

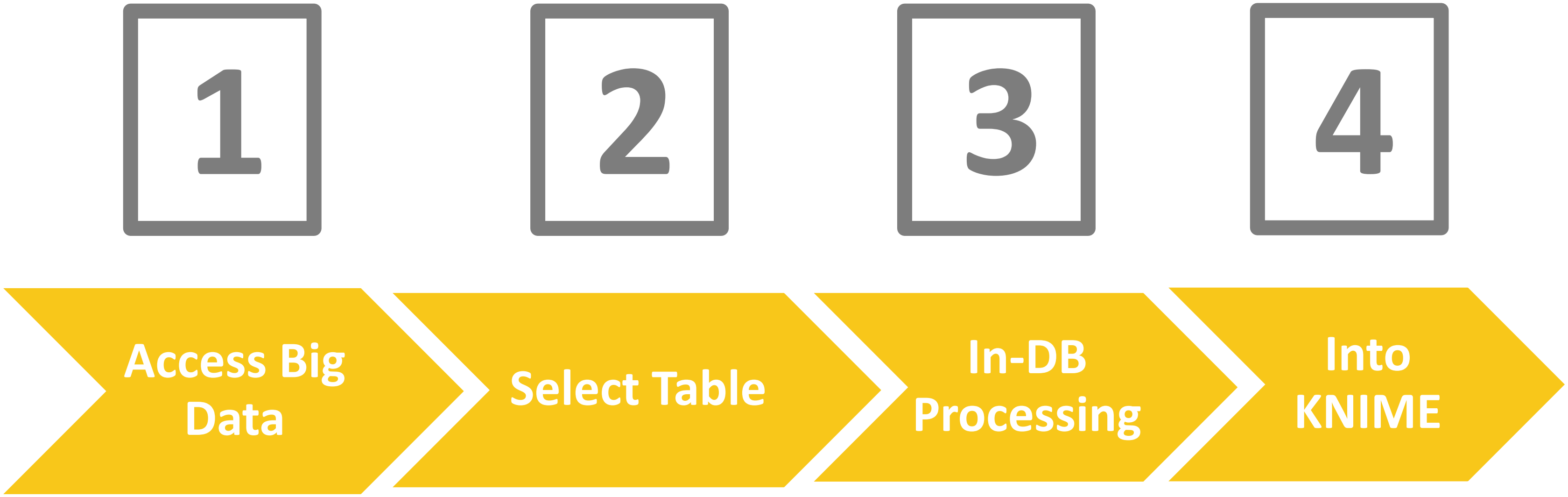
- Virtual Box

  https://www.virtualbox.org/

- VMWare Player

  http://www.vmware.com/

… as easy as 1,2,3,… 4



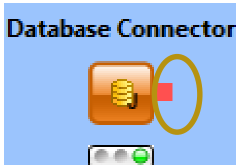
1. Access Big Data
2. Select Table
3. In-DB Processing
4. Into KNIME

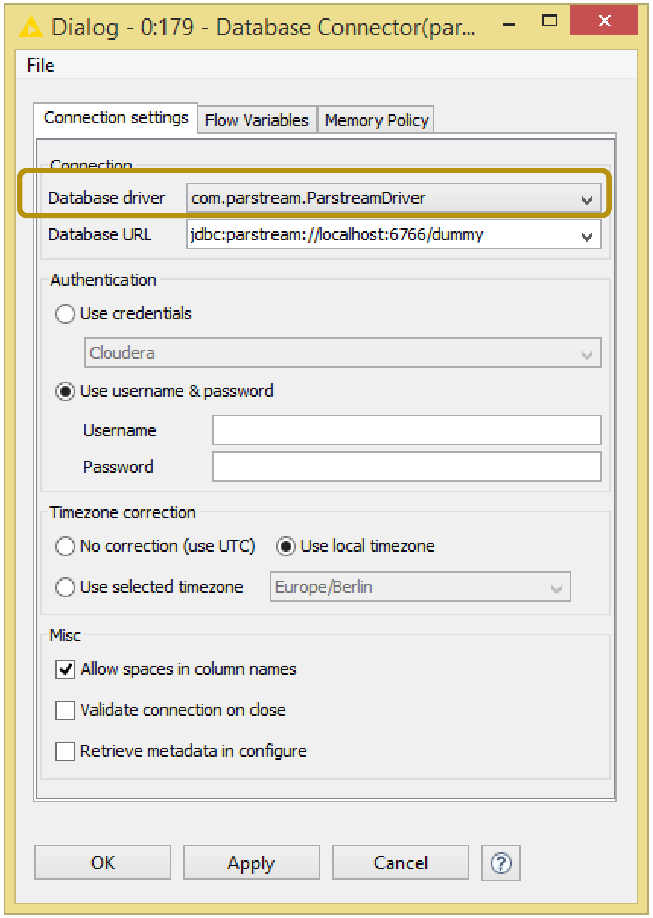
1. Database Connector

Access Big Data

Generic Database Connector

- Can connect to any JDBC source
- Register new JDBC driver via preferences page





Connection

Database driver: com.parstream.ParstreamDriver

Database URL: jdbc:parstream://localhost:6766/dummy

Authentication

○ Use credentials

Cloudera

● Use username & password

Username:

Password:

Timezone correction

○ No correction (use UTC) ● Use local timezone

○ Use selected timezone Europe/Berlin

Misc

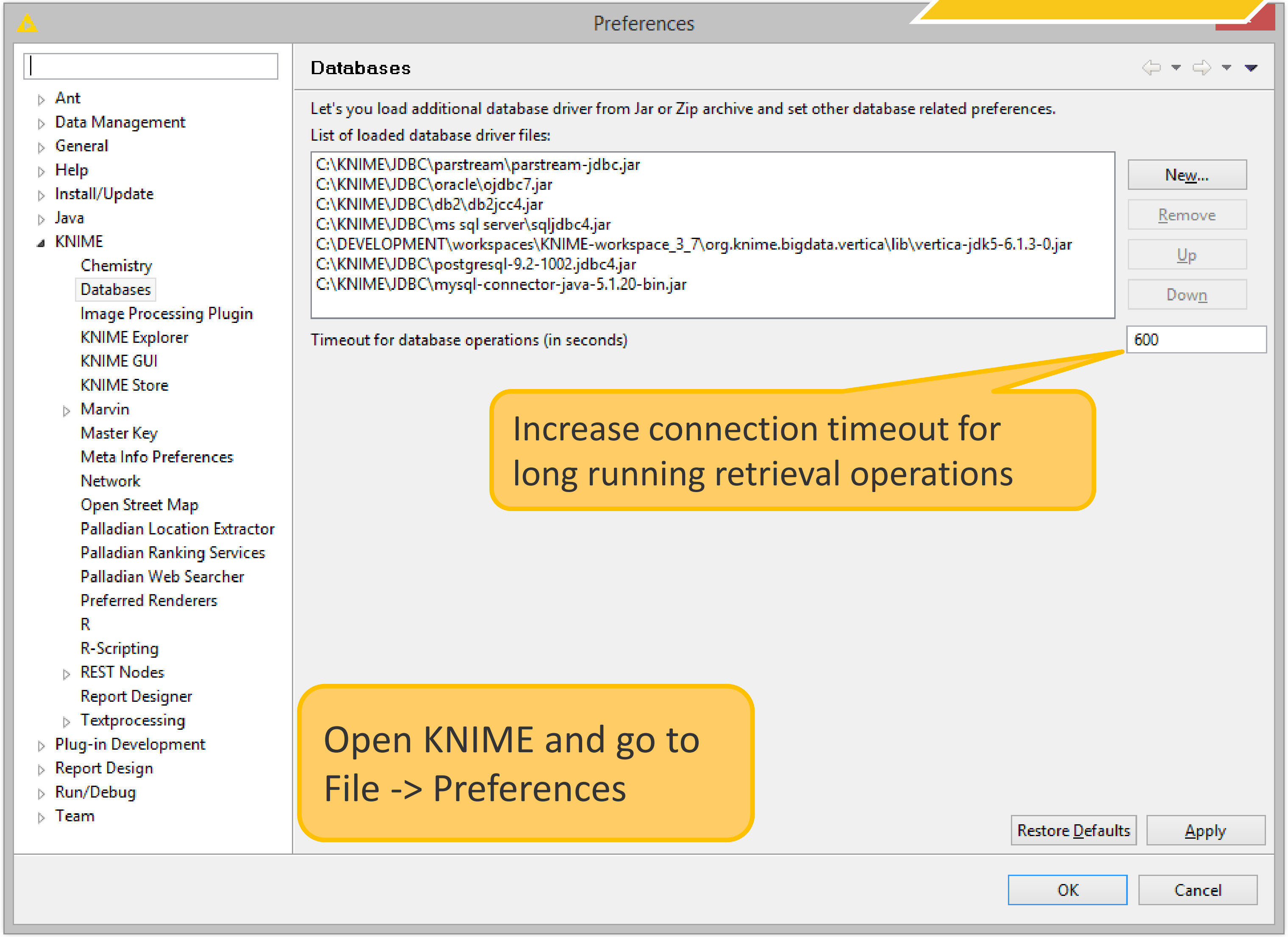
☑ Allow spaces in column names

☐ Validate connection on close

☐ Retrieve metadata in configure

1. Register JDBC Driver

Access Big Data



Databases

Let's you load additional database driver from Jar or Zip archive and set other database related preferences.

List of loaded database driver files:

```
C:\KNIME\JDBC\parstream\parstream-jdbc.jar
C:\KNIME\JDBC\oracle\ojdbc7.jar
C:\KNIME\JDBC\db2\db2jcc4.jar
C:\KNIME\JDBC\ms sql server\sqljdbc4.jar
C:\DEVELOPMENT\workspaces\KNIME-workspace_3_7\org.knime.bigdata.vertica\lib\vertica-jdk5-6.1.3-0.jar
C:\KNIME\JDBC\postgresql-9.2-1002.jdbc4.jar
C:\KNIME\JDBC\mysql-connector-java-5.1.20-bin.jar
```

Timeout for database operations (in seconds): 600

Increase connection timeout for long running retrieval operations

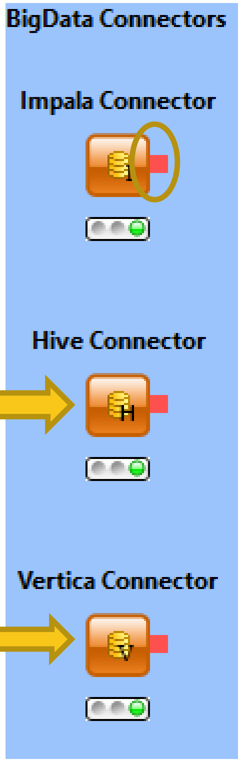
Open KNIME and go to File -> Preferences

1. Dedicated Connectors

Access Big Data

Dedicated pre-configured connectors

- Bundling necessary JDBC drivers
- Easy to use
- DB specific behavior/capability



works for most Hadoop HIVE installations, including **Hortonworks**

free

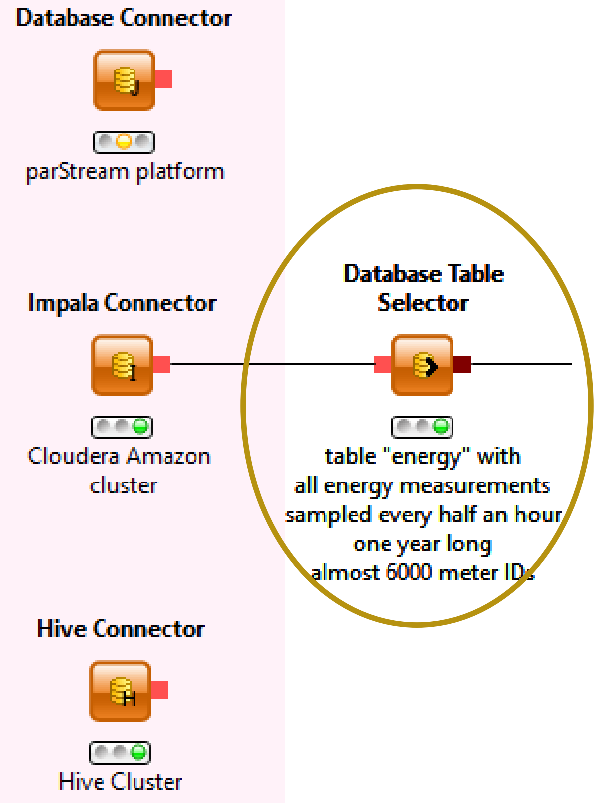
Some dedicated connectors are part of the open source KNIME Analytics Platform, some belong to the commercial KNIME Big Data Extension

2. Data Table Selection

Select Table

Connect to a big data platform:
- Impala
- Hive
- parStream

to read the energy data



Database Connector

parStream platform

Impala Connector

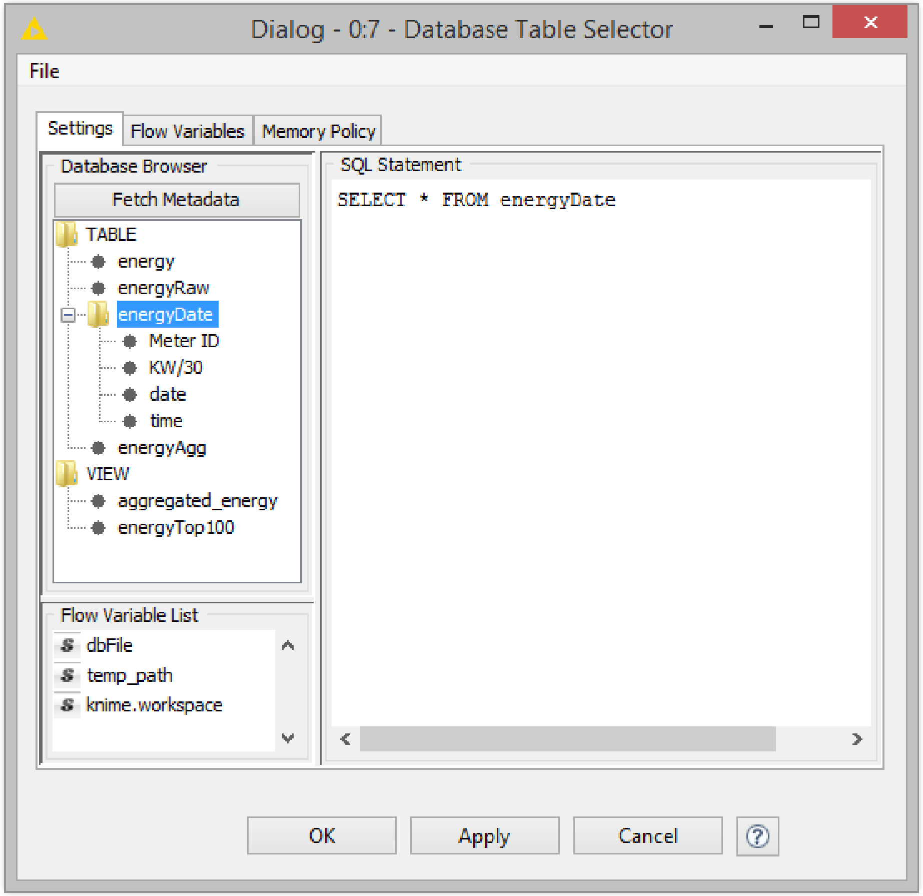
Cloudera Amazon cluster

Database Table Selector

table "energy" with all energy measurements sampled every half an hour one year long almost 6000 meter IDs

Hive Connector

Hive Cluster



Dialog - 0:7 - Database Table Selector

- TABLE
  - energy
  - energyRaw
  - energyDate
    - Meter ID
    - KW/30
    - date
    - time
  - energyAgg
- VIEW
  - aggregated_energy
  - energyTop100

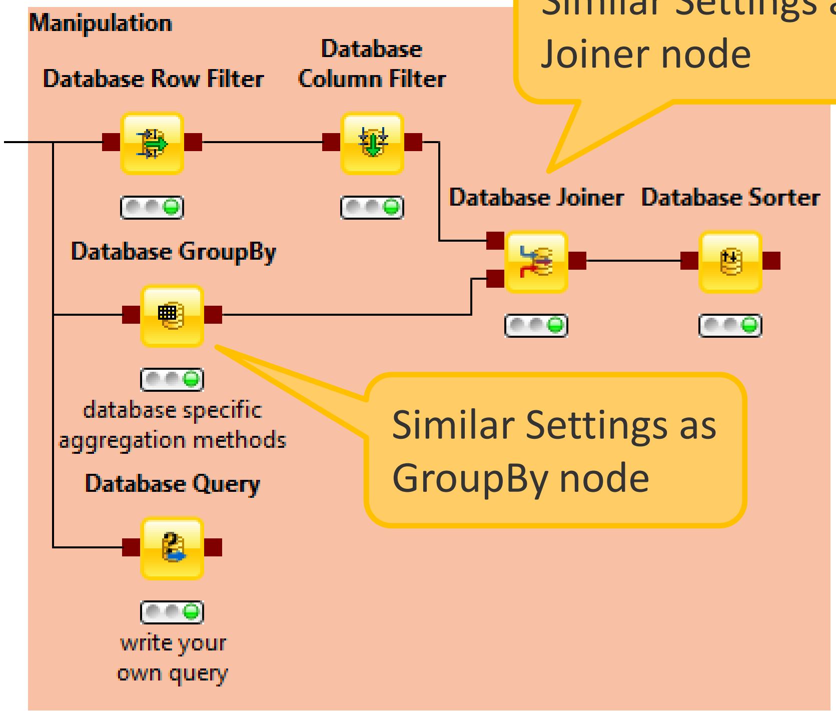
Flow Variable List: dbFile, temp_path, knime.workspace

```
SELECT * FROM energyDate
```

3. In-Database Processing

In-DB Processing

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



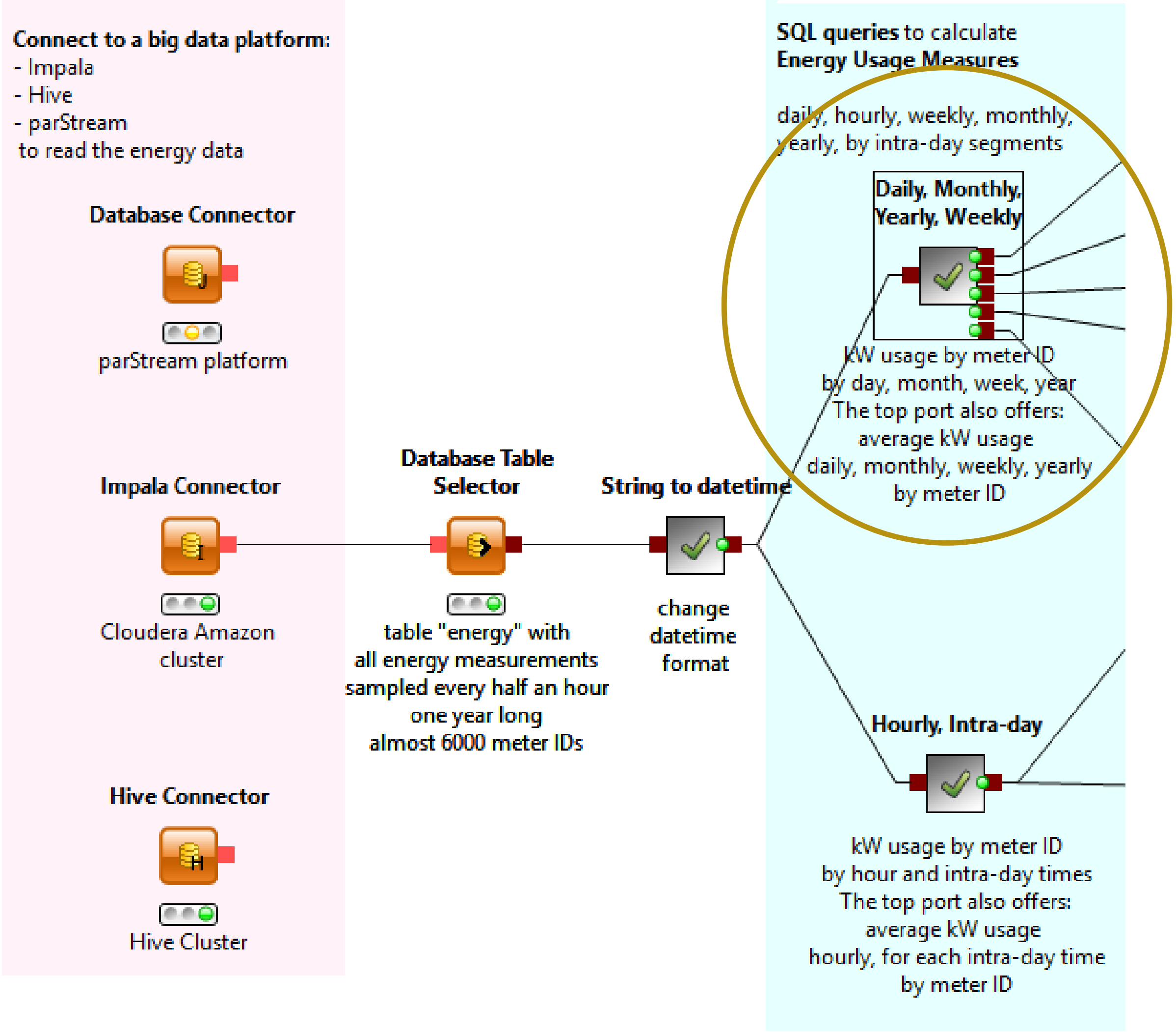
Similar Settings as Joiner node

Similar Settings as GroupBy node

* Database GroupBy node exposes DB specific aggregation methods

3. Queries for average Measures

In-DB Processing



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

String to datetime

change datetime format

SQL queries to calculate
Energy Usage Measures

daily, hourly, weekly, monthly, yearly, by intra-day segments

Daily, Monthly, Yearly, Weekly

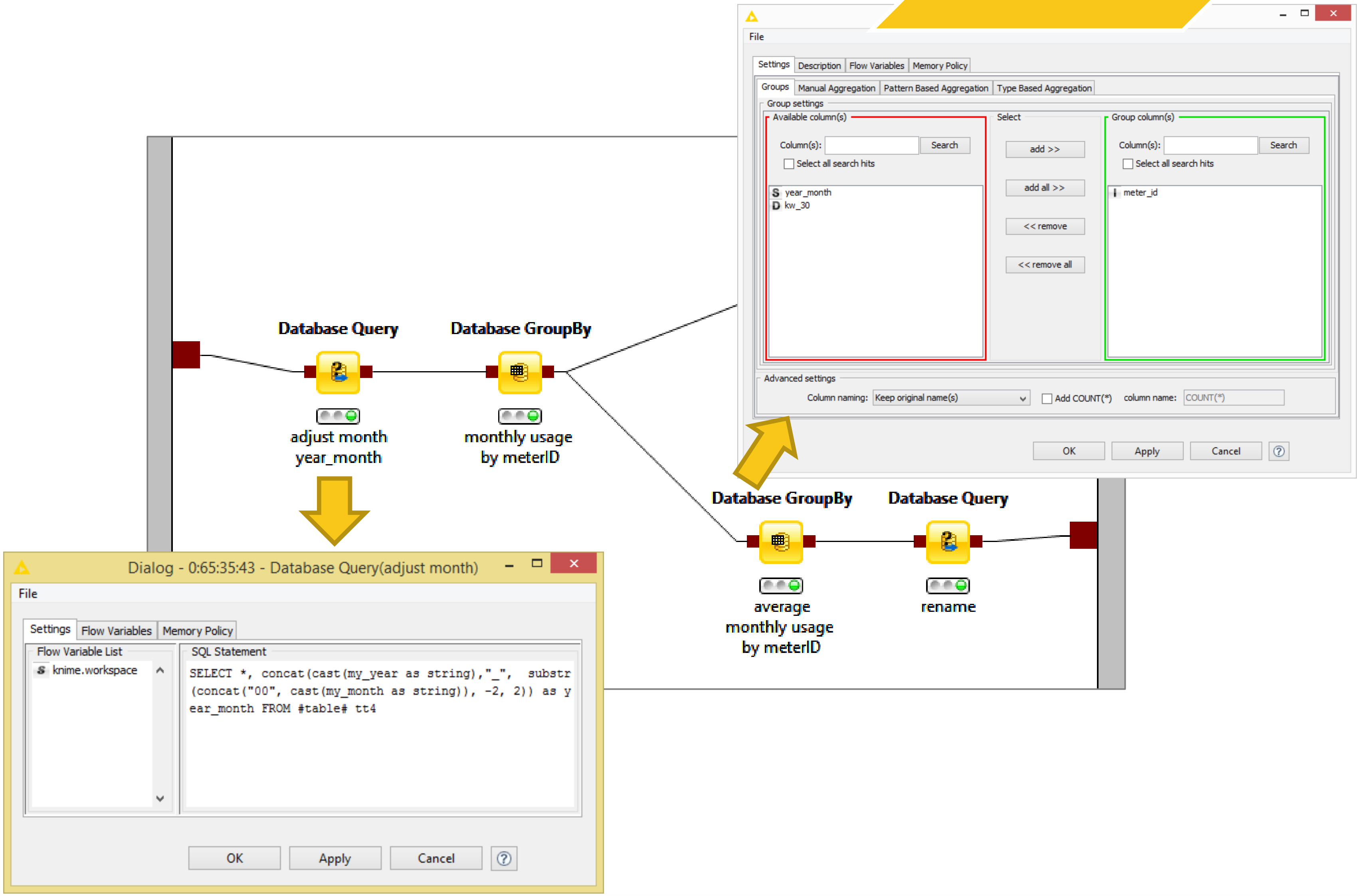
kW usage by meter ID
by day, month, week, year
The top port also offers:
average kW usage
daily, monthly, weekly, yearly
by meter ID

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

3. Average Monthly Values

In-DB Processing



Group settings

Available column(s)

Column(s):

□ Select all search hits

- S year_month
- D kw_30

Select

Group column(s)

Column(s):

□ Select all search hits

- I meter_id

Advanced settings

Column naming: Keep original name(s) □ Add COUNT(*) column name: COUNT(*)

Flow Variable List

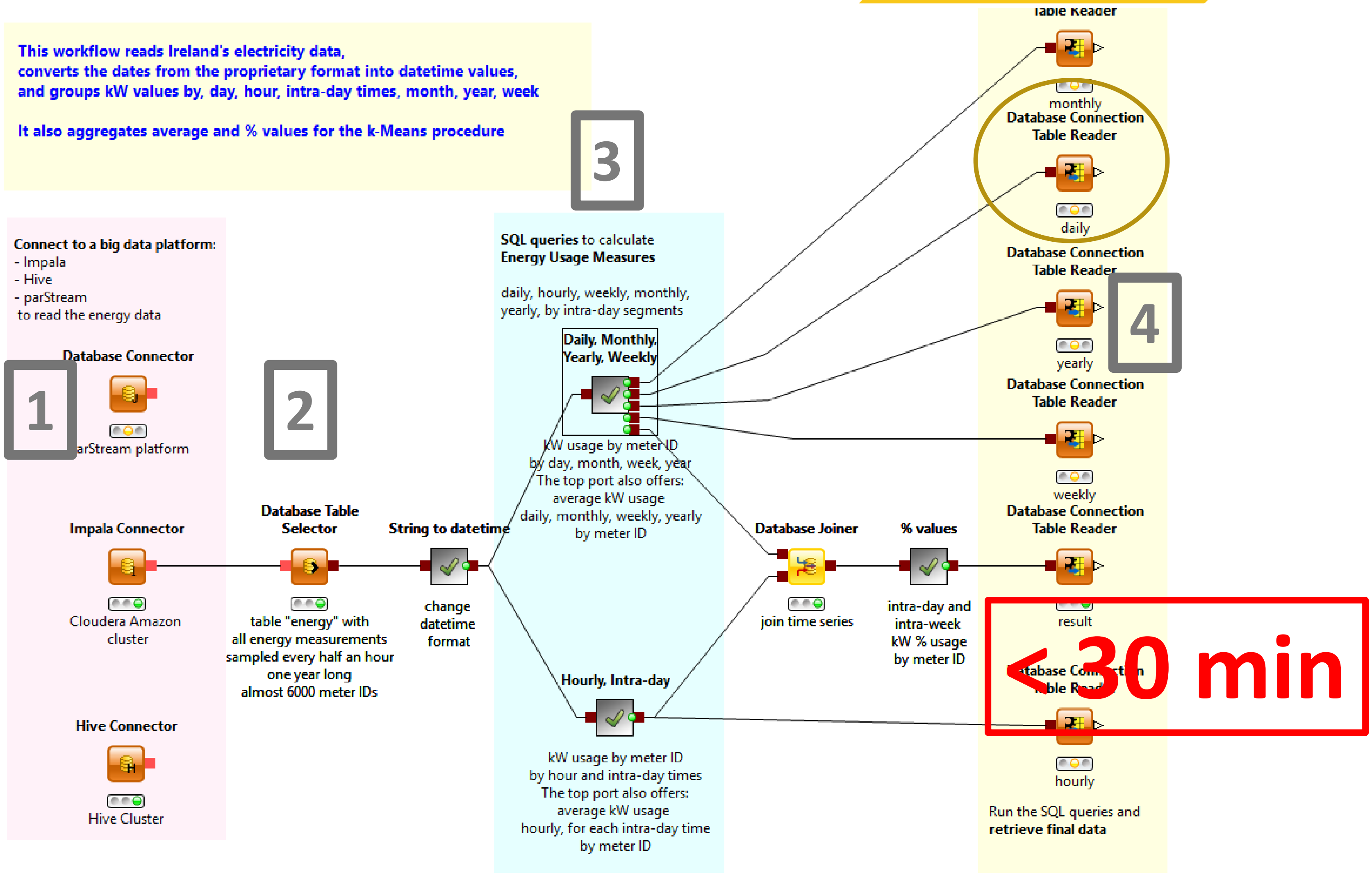
- knime.workspace

SQL Statement

```
SELECT *, concat(cast(my_year as string),"_", substr
(concat("00", cast(my_month as string)), -2, 2)) as y
ear_month FROM #table# tt4
```

4. 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

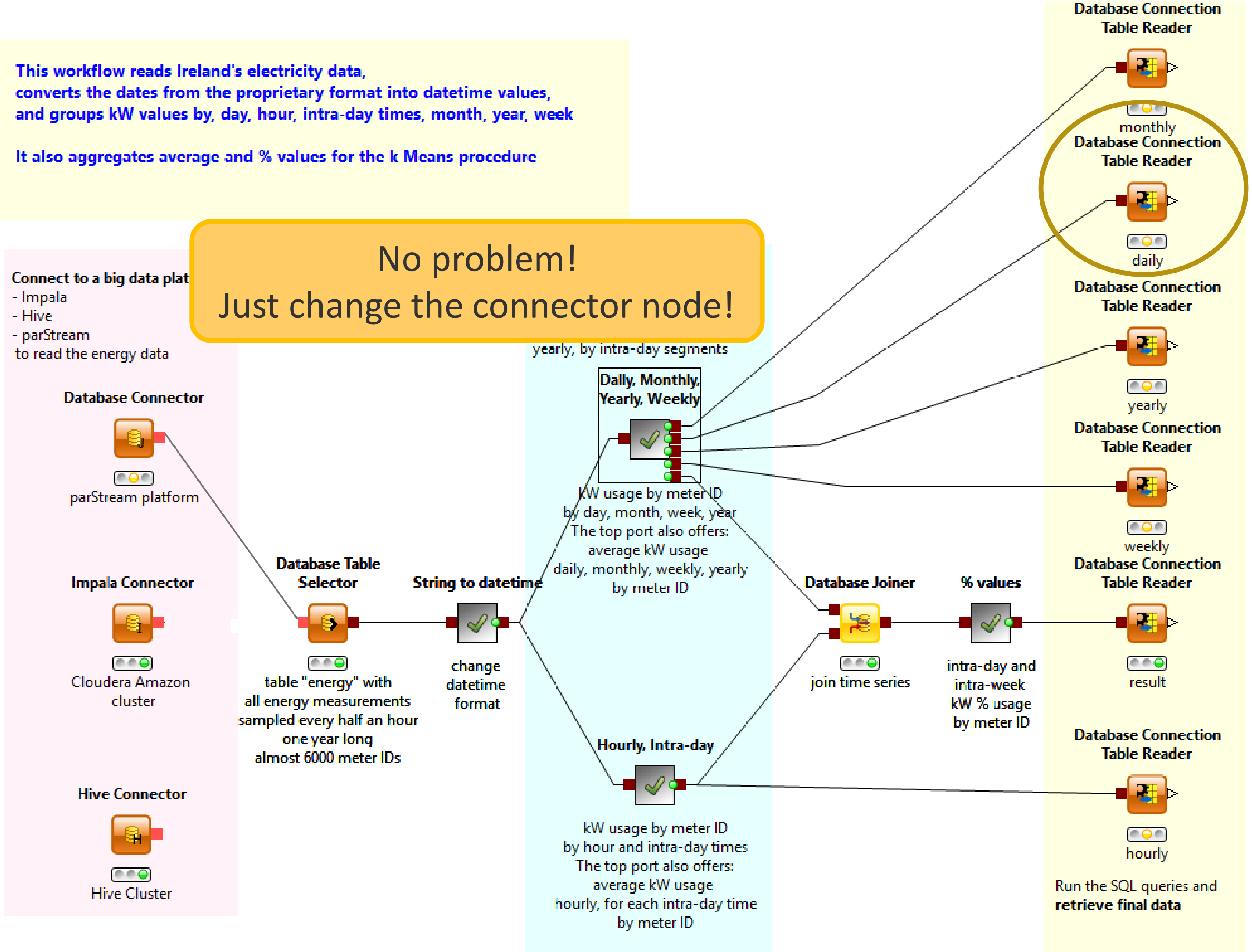
kW usage by meter ID by day, month, week, year
The top port also offers: average kW usage daily, monthly, weekly, yearly by meter ID

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

Run the SQL queries and retrieve final data

1
2
3
4
< 30 min
Database Connector
arStream 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
String to datetime
change datetime format
Daily, Monthly, Yearly, Weekly
Hourly, Intra-day
Database Joiner
join time series
% values
intra-day and intra-week kW % usage by meter ID
Table Reader
monthly
Database Connection Table Reader
daily
Database Connection Table Reader
yearly
Database Connection Table Reader
weekly
Database Connection Table Reader
result
hourly

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

Connect to a big data plat
- Impala
- Hive
- parStream
to read the energy data

yearly, by intra-day segments

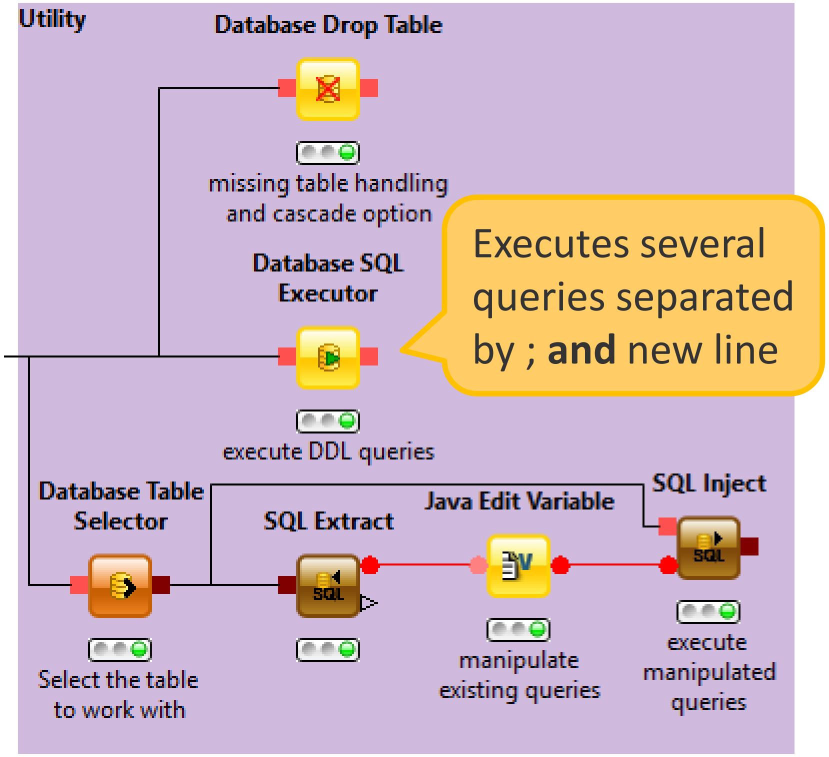
Run the SQL queries and retrieve final data

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
String to datetime
change datetime format
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 Joiner
join time series
% values
intra-day and intra-week kW % usage by meter ID
Database Connection Table Reader
monthly
Database Connection Table Reader
daily
Database Connection Table Reader
yearly
Database Connection Table Reader
weekly
Database Connection Table Reader
result
Database Connection Table Reader
hourly

No problem!
Just change the connector node!

Other Useful Database Nodes

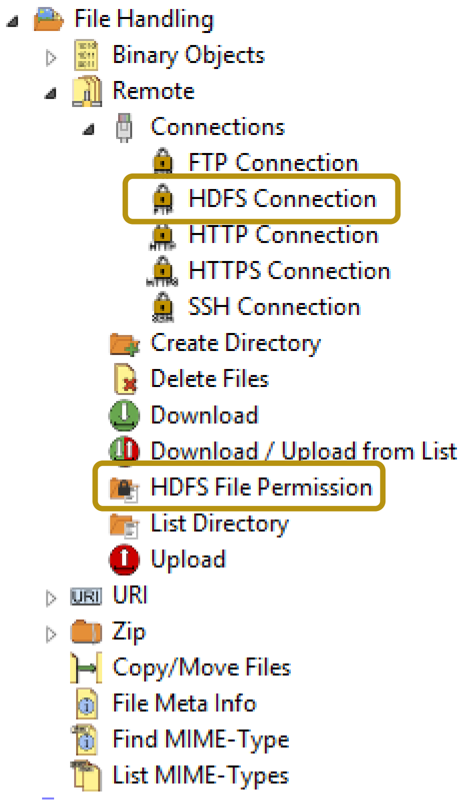
- Drop table
  - missing table handling
  - cascade option
- Execute any SQL statement
- Manipulate existing queries



Executes several queries separated by ; **and** new line

HDFS File Handling

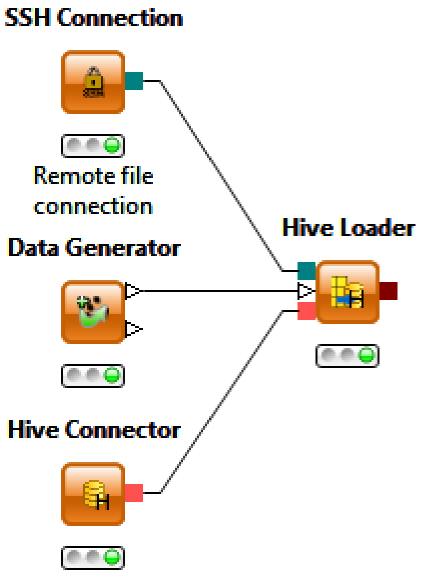
- KNIME & Extensions -> KNIME File Handling Nodes
- HDFS Connection and HDFS File Permission nodes

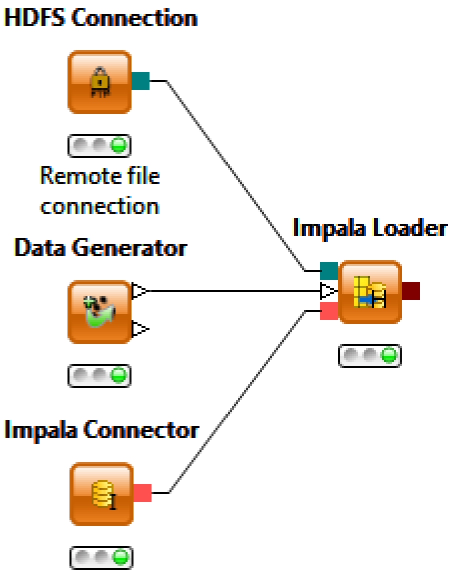


File Handling
Binary Objects
Remote
Connections
FTP Connection
HDFS Connection
HTTP Connection
HTTPS Connection
SSH Connection
Create Directory
Delete Files
Download
Download / Upload from List
HDFS File Permission
List Directory
Upload
URI
Zip
Copy/Move Files
File Meta Info
Find MIME-Type
List MIME-Types

Hive/Impala Loader

- Upload a KNIME data table to Hive/Impala

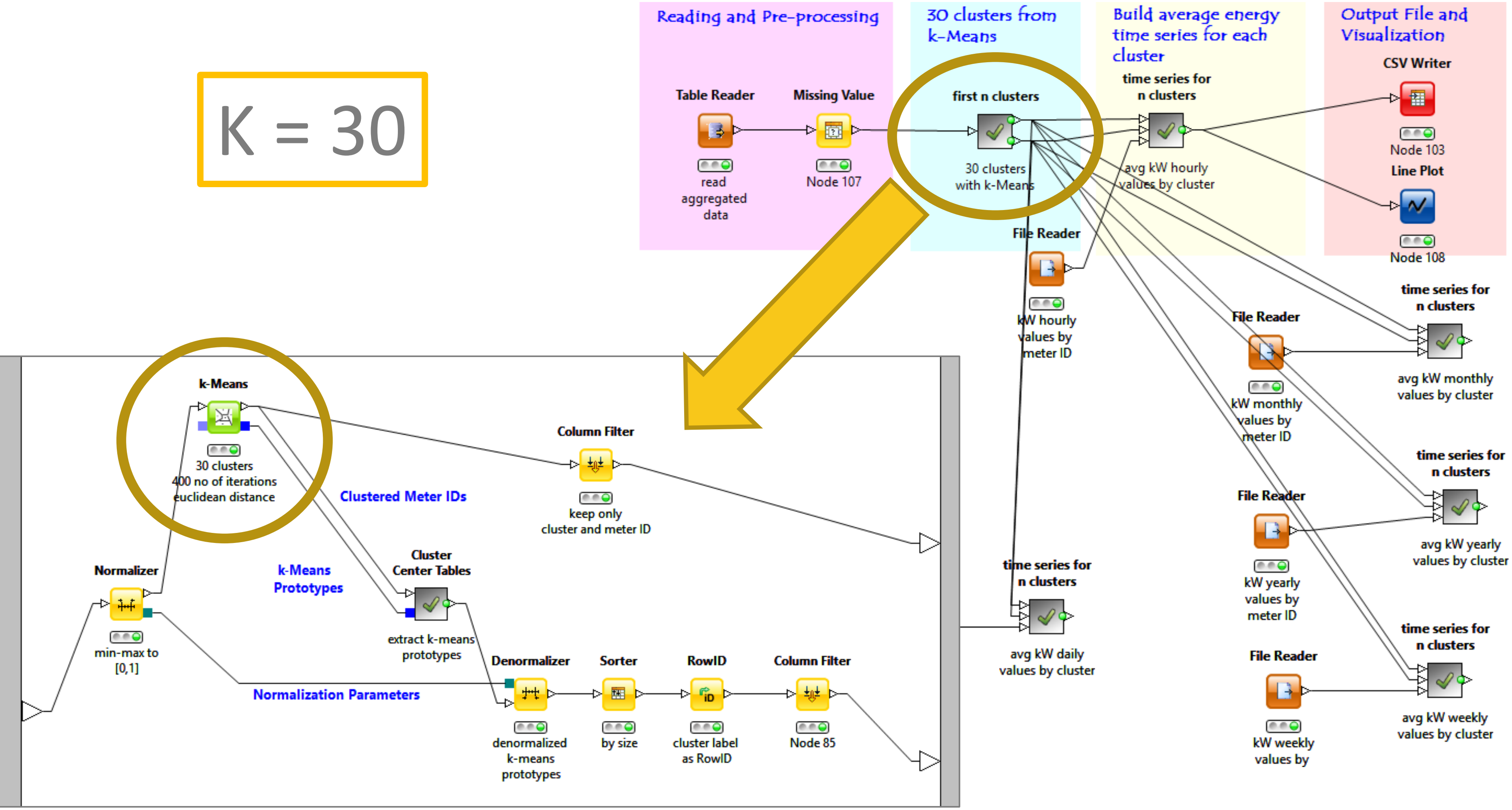




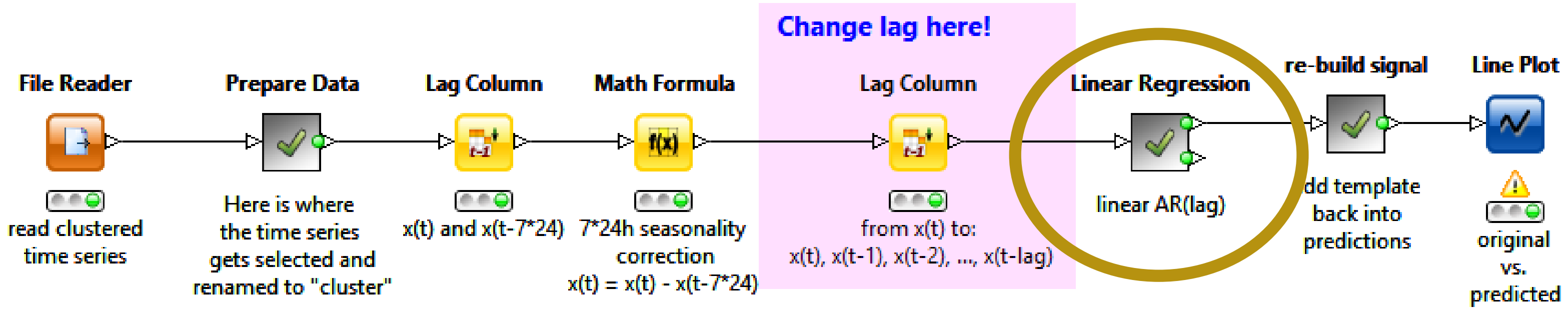
Next Steps

Following Workflows: k-Means

K = 30

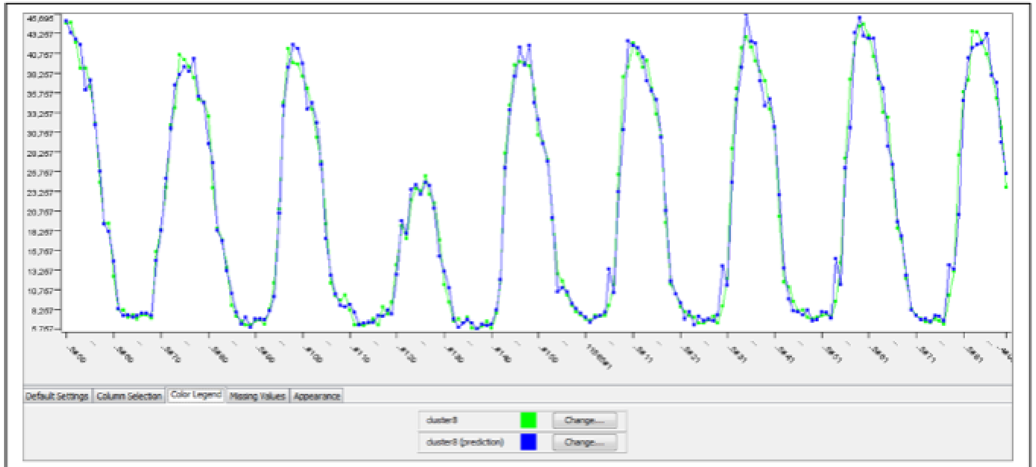

Reading and Pre-processing
Table Reader
read aggregated data
Missing Value
Node 107
30 clusters from k-Means
first n clusters
30 clusters with k-Means
Build average energy time series for each cluster
time series for n clusters
avg kW hourly values by cluster
Output File and Visualization
CSV Writer
Node 103
Line Plot
Node 108
File Reader
kW hourly values by meter ID
time series for n clusters
avg kW daily values by cluster
File Reader
kW monthly values by meter ID
time series for n clusters
avg kW monthly values by cluster
File Reader
kW yearly values by meter ID
time series for n clusters
avg kW yearly values by cluster
File Reader
kW weekly values by
time series for n clusters
avg kW weekly values by cluster
k-Means
30 clusters
400 no of iterations
euclidean distance
Clustered Meter IDs
Column Filter
keep only cluster and meter ID
Normalizer
min-max to [0,1]
k-Means Prototypes
Cluster Center Tables
extract k-means prototypes
Normalization Parameters
Denormalizer
denormalized k-means prototypes
Sorter
by size
RowID
cluster label as RowID
Column Filter
Node 85

Following Workflows: AR Model



*Auto-regressive model using the previous 24h*7 as seasonality template*

- **24-hour seasonality template:** the first week of the time series is used as a template for seasonality correction
- **auto** means usage of past of the same time series for prediction. No other external time series/data used.
- **Regressive** refers to the mode used: either a linear or a polynomial regression



cluster8 🟩 Change...
cluster8 (prediction) 🟦 Change...

Figure 30: Original time series and predicted time series after being adjusted for weekly seasonality in green and blue respectively.

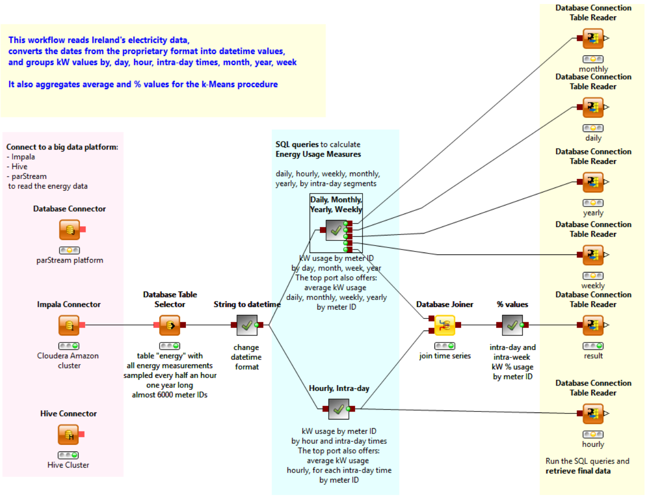
Model Factory: Concept

Data Preparation on

Hadoop

K-Means

A-R Model



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

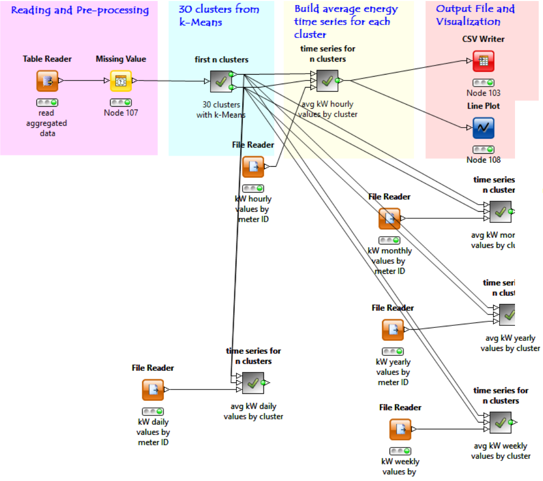
Connect to a big data platform:
- Impala
- Hive
- parStream
to read the energy data

SQL queries to calculate **Energy Usage Measures**

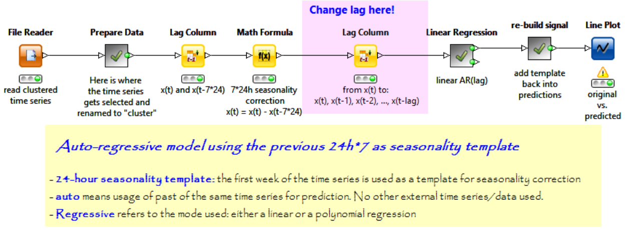
daily, hourly, weekly, monthly, yearly, by intra-day segments

Run the SQL queries and **retrieve final 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
String to datetime
change datetime format
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 Joiner
join time series
% values
intra-day and intra-week kW % usage by meter ID
Database Connection Table Reader
monthly
Database Connection Table Reader
daily
Database Connection Table Reader
yearly
Database Connection Table Reader
weekly
Database Connection Table Reader
result
Database Connection Table Reader
hourly



Reading and Pre-processing
30 clusters from k-Means
Build average energy time series for each cluster
Output File and Visualization
Table Reader
read aggregated data
Missing Value
Node 107
first n clusters
30 clusters with k-Means
time series for n clusters
avg kW hourly values by cluster
CSV Writer
Node 103
Line Plot
Node 108
File Reader
kW hourly values by meter ID
File Reader
kW monthly values by meter ID
time series n cluster
avg kW mor values by clu
File Reader
kW yearly values by meter ID
time seri n clust
avg kW yearly values by cluster
File Reader
kW daily values by meter ID
time series for n clusters
avg kW daily values by cluster
File Reader
kW weekly values by
time series for n clusters
avg kW weekly values by cluster



*Auto-regressive model using the previous 24h*7 as seasonality template*

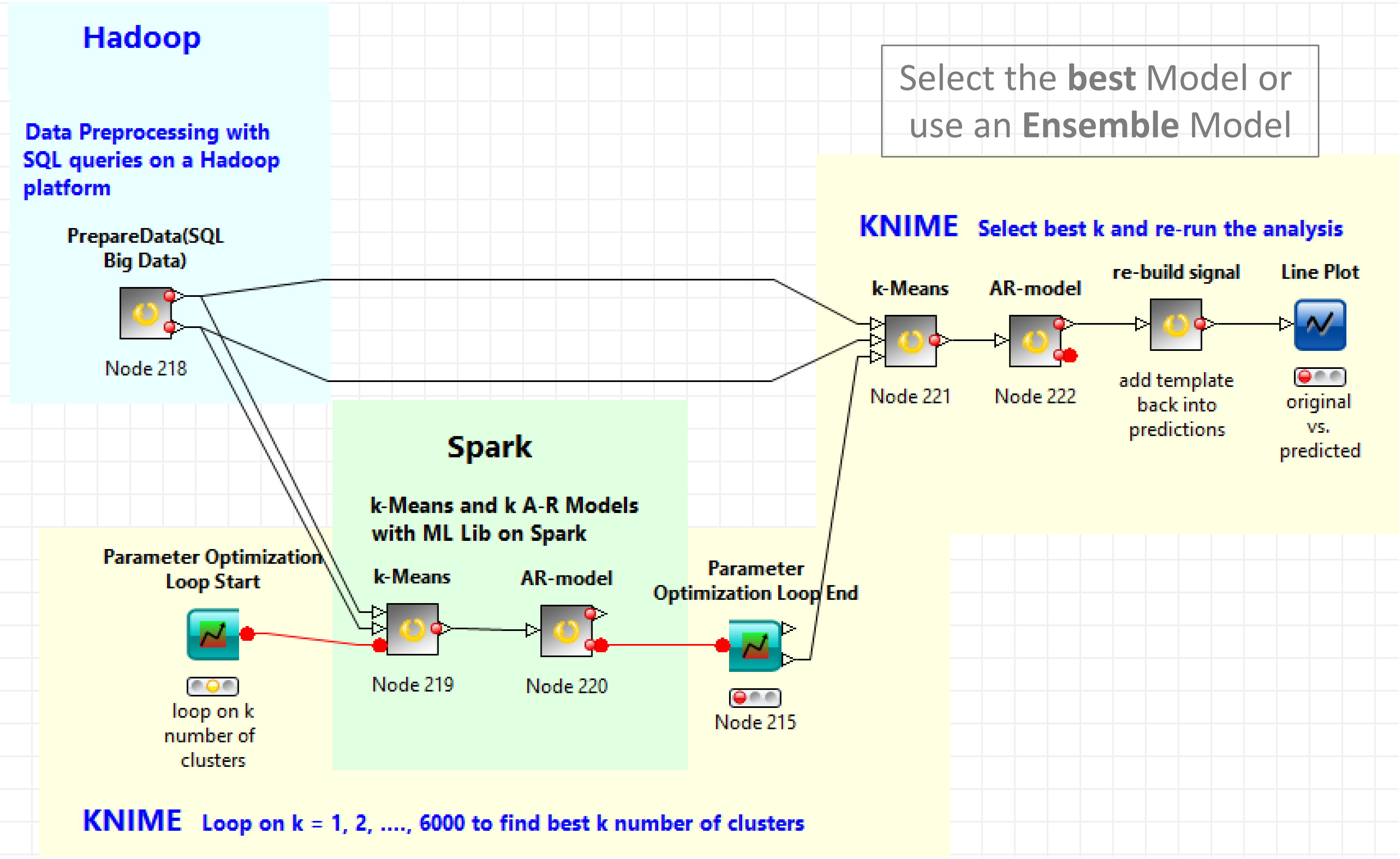
- **24-hour seasonality template:** the first week of the time series is used as a template for seasonality correction
- **auto** means usage of past of the same time series for prediction. No other external time series/data used.
- **Regressive** refers to the mode used: either a linear or a polynomial regression

Change lag here!
File Reader
read clustered time series
Prepare Data
Here is where the time series gets selected and renamed to "cluster"
Lag Column
x(t) and x(t-7*24)
Math Formula
7*24h seasonality correction x(t) = x(t) - x(t-7*24)
Lag Column
from x(t) to: x(t), x(t-1), x(t-2), ..., x(t-lag)
Linear Regression
linear AR(lag)
re-build signal
add template back into predictions
Line Plot
original vs. predicted

Choosing best k for k-Means minimizing the RMS prediction error from A-R Model on **Spark**

Loop, control, and final results remain in **KNIME**

Model Factory: Workflow



Select the **best** Model or use an **Ensemble** Model

Hadoop

Data Preprocessing with SQL queries on a Hadoop platform

PrepareData(SQL Big Data)

Node 218

Spark

k-Means and k A-R Models with ML Lib on Spark

k-Means

Node 219

AR-model

Node 220

Parameter Optimization Loop Start

loop on k number of clusters

Parameter Optimization Loop End

Node 215

KNIME Loop on k = 1, 2, ..., 6000 to find best k number of clusters

KNIME Select best k and re-run the analysis

k-Means

Node 221

AR-model

Node 222

re-build signal

add template back into predictions

Line Plot

original vs. predicted

References

- Whitepaper “KNIME opens the Doors to Big Data”

  http://www.knime.org/files/big_data_in_knime_1.pdf

- 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

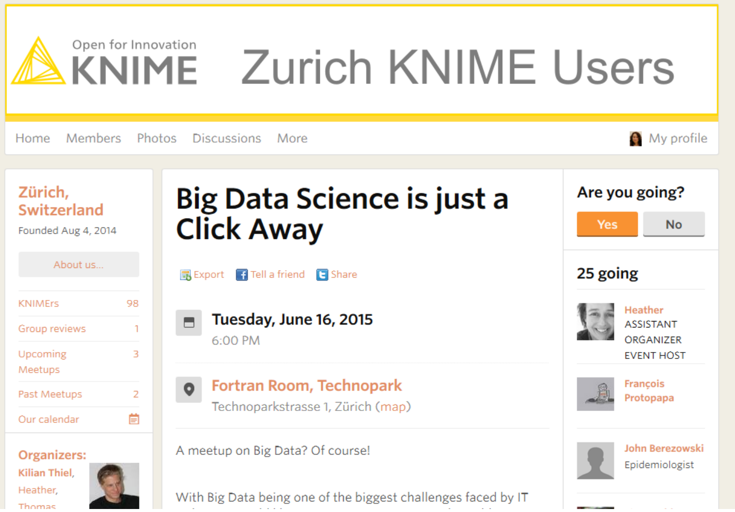
- The Big Data Extension Product Description
  http://www.knime.org/knime-big-data-extension

Resources

- KNIME (www.knime.org)
  - BLOG for news, tips and tricks(www.knime.org/blog)
  - FORUM for questions and answers (tech.knime.org/forum)
  - EXAMPLE SERVER for example workflows
  - LEARNING HUB (www.knime.org/learning-hub)
- KNIME TV channel on YouTube
- KNIME on 🐦 @KNIME
- KNIME on f https://www.facebook.com/KNIMEanalytics

Events and Trainings

- **KNIME** (https://www.knime.org/about/events)
  - **User Training 15-16 June** Zurich (https://www.knime.org/knime-user-training-june-2015)
  - **Developer Training 17-18 June** Zurich (https://www.knime.org/knime-developer-training-june-2015)
- **Meetup.com**
  - **16 June**, Zurich, Meetup on **Big Data** (http://www.meetup.com/Zurich-KNIME-Users/events/221570664/)



Open for Innovation KNIME Zurich KNIME Users

Zürich, Switzerland
Founded Aug 4, 2014

KNIMErs	98
Group reviews	1
Upcoming Meetups	3
Past Meetups	2
Our calendar	

Organizers: Kilian Thiel, Heather, Thomas

Big Data Science is just a Click Away

Tuesday, June 16, 2015
6:00 PM

Fortran Room, Technopark
Technoparkstrasse 1, Zürich (map)

A meetup on Big Data? Of course!

With Big Data being one of the biggest challenges faced by IT

Are you going?

25 going

Heather
ASSISTANT ORGANIZER
EVENT HOST

François Protopapa

John Berezowski
Epidemiologist

Thank you!