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...

- 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
IoT

THE INTERNET OF THINGS
AN EXPLOSION OF CONNECTED POSSIBILITY

BILLIONS OF DEVICES

2002 1,000,000
2008 0.5 BILLION
2009 IoT INCEPTION
2012 8.7 BILLION
2013 11.2 BILLION
2014 14.4 BILLION
2016 22.9 BILLION
2018 34.8 BILLION
2019 42.1 BILLION
2020 50.1 BILLION

Copyright © 2015 KNIME.com AG
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

Joiner
% values

Write to server
Write to CSV

~ 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)
Hadoop Sandboxes

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

• Cloudera:

• 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

Generic Database Connector

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

Open KNIME and go to File -> Preferences

Access Big Data

Increase connection timeout for long running retrieval operations
1. Dedicated Connectors

Dedicated pre-configured connectors

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

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

works for most Hadoop HIVE installations, including Hortonworks

free
2. 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.

Dialog - 0:7 - Database Table Selector

File

Settings Flow Variables Memory Policy

Database Browser

Fetch Metadata

TABLE

- energy
- energyRaw
- energyDate

- Meter ID
- KW/30
- date
- time
- energyApp
- aggregated_energy
- energyTop100

VIEW

- dbFile
- temp_path
- knime.workspace

SQL Statement

SELECT * FROM energyDate
3. In-Database Processing

- 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
3. 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
3. Average Monthly Values

In-DB Processing

Database Query

- Adjust month
- Year month

Database GroupBy

- Monthly usage
- By meterID

Dialog - 06:35:43 - Database Query (adjust month)

- SQL Statement:
  ```sql
  SELECT *, concat(cast(my_year as string), " ", substr(
    concat("0", cast(my_month as string)), -2, 2) as y
  from month_usage_flows.
  ```

Database GroupBy

- Average
- Monthly usage
- By meterID

Database Query

- Rename
4. Import Data from Database

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.

1. Connect to a big data platform:
   - Impala
   - Hive
   - prestoStream to read the energy data

2. Database connectors:
   - Impala Connector
   - Hive Connector
   - Cloudere Amazon cluster

3. SQL queries to calculate Energy Usage Measures:
   - 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 intra-day segments
   - 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

4. Run the SQL queries and retrieve final data
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.

New Big Data Platform?

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
Hive/Impala Loader

- Upload a KNIME data table to Hive/Impala
Next Steps
Following Workflows: k-Means

K = 30
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

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

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
Select the best Model or use an Ensemble Model
References

• 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

• The Big Data Extension Product Description
  http://www.knime.org/knime-big-data-extension
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)
- **KNIME on** [Twitter](https://twitter.com) @KNIME
- **KNIME on** [Facebook](https://www.facebook.com) https://www.facebook.com/KNIMEanalytics
Events and Trainings

• **KNIME** ([https://www.knime.org/about/events](https://www.knime.org/about/events))
• **Meetup.com**

Thank you!