Instrumenting KNIME with KNIME
(or how the shoemaker eats her own boots)

Iris Adä
KNIME.com AG
WITHOUT DATA YOU ARE JUST ANOTHER PERSON WITH AN OPINION

*W. Edwards Deming*
Will They Blend? blog post series

Table of content
- SAS, SPSS, and MATLAB meet Amazon S3:
- XML meets JSON
- MS Access meets H2. Test your Baseball Kn
- Amazon S3 meets MS Azure Blob Storage. A
- Local vs. remote files. Will blending overcom
- MS Word meets Web Crawling. Identifying th
- A Cross-Platform Ensemble Model: R meets
- Twitter meets PostgreSQL. More than idle ch
- Hadoop Hive meets Excel. Your flight is boa
- Open Street Maps (OSM) meets CSV Files a
- IBM Watson meets Google API

https://www.knime.org/blog/will-they-blend
The Goal

• Answer questions!
• One report with numbers from different sources in a unique/comparable way
• Four dimensions:
  – Per week, per month, per quarter, per year
• Fully automated

Here:
Will they be read?
The report
The Report - Weeks

Most recent indicators
Compare weekly aggregates
Inspect the last two weeks daily
The Report - Months

Monthly aggregates

Compare monthly aggregates over years

Inspect the last two month daily
The Report - Quarters

Most recent indicators

Compare quarterly aggregates

Inspect the last two weeks daily
The Report - Years

Most recent indicators

Compare quarterly aggregates

Inspect the last two weeks daily
From data to report

Get the data and aggregate it

Generate the report

Deploy! (Execute and send the report)
Find the blogs

<table>
<thead>
<tr>
<th>Table Creator</th>
<th>HttpRetriever</th>
<th>HtmlParser</th>
<th>XPath</th>
<th>Ungroup</th>
<th>XPath</th>
<th>Filter duplicates</th>
<th>Row Splitter</th>
<th>Concatenate</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.knime.org/blog">www.knime.org/blog</a></td>
<td>Extract all links and tags</td>
<td>Extract href attribute</td>
<td>Pagination URLs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Column Rename</td>
<td>HttpRetriever</td>
<td>HtmlParser</td>
<td>XPath</td>
<td>Ungroup</td>
<td>XPath</td>
<td>Row Filter</td>
<td>Column Filter</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Extract all links and tags</td>
<td>Extract href attribute</td>
<td>Only blog URLs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>String Manipulation</td>
<td>HttpRetriever</td>
<td>HtmlParser</td>
<td>XPath</td>
<td>Column Filter</td>
<td>Create search filters</td>
<td>Column Filter</td>
<td>String Manipulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Extract all links H1 tags as titles</td>
<td></td>
<td></td>
<td></td>
<td>Create search filters</td>
<td>Column Resorter</td>
<td>Column Filter</td>
<td>String Manipulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Concatenate</td>
</tr>
</tbody>
</table>

Get the data and aggregate it

Deploy! (Execute and send the report)
<table>
<thead>
<tr>
<th>Search Filter</th>
<th>Search Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>/blog(?!:|||?+)?</td>
<td>Blog TOC Page</td>
</tr>
<tr>
<td>/blog/?(?!:|||?+)?</td>
<td>Blog (any)</td>
</tr>
<tr>
<td>/blog/?creating-restful-services-with-knime-an-update</td>
<td>Creating RESTful services with KNIME, an update</td>
</tr>
</tbody>
</table>
| /blog/?creating-restful-services-with-knime-an-update(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: SAS, SPSS, and MATLAB meet Amazon S3: setting the first wave of data science innovation
| /blog/?SAS-SPSS-Matlab-meet-S3(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: XML meets JSON |
| /blog/?XML-meets-JSON(?!:\|\|\|\?\+)? | Will they blend? Experiments in Data & Tool Blending: Today: IBM Watson meets Google API |
| /blog/?IBM-Watson-meets-Google-API(?!:\|\|\|\?\+)? | Using custom data types with the Java Snippet node |
| /blog/?MSAccess_meets_H2(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: Amazon S3 meets MS Azure Blob Storage. A match made in heaven
| /blog/?S3_meets_BlobStorage(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: Local vs. remote files. Will blending overcome the distance
| /blog/?local-ZIP-meets-remote-ZIP(?!:\|\|\|\?\+)? | Optimizing KNIME workflows for performance |
| /blog/?optimizing-knime-workflows-for-performance(?!:\|\|\|\?\+)? | Speaking Kerberos with KNIME Big Data Extensions |
| /blog/?speaking-kerberos-with-knime-big-data-extensions(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: MS Word meets Web Crawling. Identifying the Secret to Success
| /blog/?MSWord_webCrawling(?!:\|\|\|\?\+)? | Will They Blend? Experiments in Data & Tool Blending: Today: A Cross-Platform Ensemble Model: R meets Python and KNIME Analytics Platform and the Web
| /blog/?KNIMEAnalyticsPlatform-meets-R-and-Python(?!:\|\|\|\?\+)? | The Node Guide, Finding Help in KNIME Analytics Platform and on the Web |
| /blog/?the-node-guide-finding-help-in-knime-analytics-platform-and-on-the-web(?!:\|\|\|\?\+)? | Will they blend? Experiments in Data & Tool Blending: Today: Twitter meets PostgreSQL. More than idle chat
| /blog/?Twitter-meets-PostgreSQL(?!:\|\|\|\?\+)? | Will they blend? Experiments in Data & Tool Blending: Today: Hadoop Hive meets Excel. Your flight is boarding now
| /blog/?Hadoop-Hive-meets-Excel(?!:\|\|\|\?\+)? | Will they blend? Experiments in Data & Tool Blending: Today: Open Street Maps (OSM) meets CSV Files and Google Just Add Imagination...
| /blog/?OSM-meets-CSV-file-and-Google-API(?!:\|\|\|\?\+)? | Integrating One More Data Source: The Semantic Web |
| /blog/?just-add-imagination(?!:\|\|\|\?\+)? | Integrating One More Data Source: The Semantic Web |
| /blog/?just-add-imagination? | 10 years of KNIME, 10 years of innovation |
| /blog/?10-years-of-knime-10-years-of-innovation(?!:\|\|\|\?\+)? | Customer Segmentation comfortably from a web browser:Until... |
| /blog/?customer-segmentation-comfortably-from-a-web-browser(?!:\|\|\|\?\+)? | 10 years of KNIME, 10 years of innovation |
Aggregation

Get the data and aggregate it

Step 1: Filter the weblog to contain only blog informations.

- Row Filter
- Table Row to Variable
- Read blogs only
- IP Blacklist
- GroupBy
- Column Rename
- Table Row to Variable

Step 2: Label each blog post with name and author

- Parallel Chunk Start
- Blog patterns
- Parallel Chunk End

Step 3: Aggregate

- String Input
- Aggregate and write

Deploy! (Execute and send the report)
Aggregation

- aggregated_snapshots
  - daily
  - general
  - monthly
    - blogs_all.table
    - forum.table
    - twitter.table
    - whitepaper_all.table
  - quarterly
  - totals
  - weekly
    - blogs_all.table
    - forum.table
    - twitter.table
    - whitepaper_all.table
  - yearly

Get the data and aggregate it

Deploy! (Execute and send the report)
Generate the report

• Goto KNIME!
This is the only workflow scheduled on the KNIME Server, once a week.
Deploy!
(Execute and send the report)
Instrumenting KNIME with KNIME
(or how the shoemaker eats her own boots)

Iris Adä
KNIME.com AG