

Kettle Past - Present - Future

Matt Casters

*Kettle Project Founder, Neo4j Chief Solutions Architect,
Project Hop Lead Architect, Crazy Belgium, Dad*



@mattcasters





The Past

is a present to the future



The past is behind us

- Kettle Open Sourced in 2005, joined Pentaho early 2006
- Thriving period
- Stabilisation period from 2014
- Acquisition by Hitachi Vantara in 2015-2016
 - Big Data
 - Hitachi technology (worker nodes)
 - Spark EE
 - Front end features
 - Enterprise Repository and security features



The recent past



Kettle/PDI release 8.2

- Python Executor Step
- Access to HCP from PDI
- Spark, AMQP Improvements
- Push-based Streaming for Dashboards
- OpenJDK support (!)
- New Data Lineage Analyzers
- Carte L&F improvements (but broken)
- Small improvements (and regressions)

Kettle/PDI release 8.3

- Snowflake support!
- Redshift Bulk loader
- Amazon Kinesis Consumer and Producer
- Hitachi Content Platform access
- AEL improvements (Switch/Case and Merge Rows support)
- Other minor step improvements and data lineage tweaks

Kettle/PDI release 9.0

- Announced for first half of 2020
- Ask Jens for details!

Plugin Machine Intelligence

- One of the coolest things ever in ML
- Welcome Mark Hall and Ken Wood
- Presentation after the coffee!

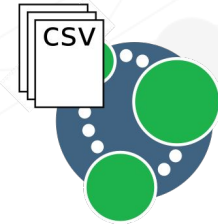


Improvements

- Data Sets plugin: CSV data sets, bug fixes, usability improvements, auto-loading, ...
- Kettle Environments plugin: easy switching, bug fixes, better handling, ...
- Needful Things plugin: more bug fixes for Kettle, Repeat job entry, ...

Improvements

- [Neo4j plugin](#)
 - A dozen or so more releases
 - Major performance improvements
 - Bulk loading support
 - Graph data type
 - [Kettle Neo4j Integration](#) project
 - Support for the Enterprise Repository



Kettle Neo4j Logging

- Write logging to Neo4j into an execution lineage and metadata graph
- Execution details are stored on Job, Job entry, Transformation, Steps and Database levels
- Documents graph node label usage by Kettle
 - Node or Relationship
 - Creation or update

New stuff - Kettle Neo4j Logging



Topic: Execution history of step '(:Link)-[:CONNECTS]-(:NE)'
Parent: 240-load-topo-links-nodes (TRANS)

#	Name	Type	R/W/I/O/R	errors	date	duration
1	(:Link)-[:CONNECTS]-(:NE)	STEP	1378/1378/0/2756/0	0	2019/01/17 14:48:16	
▼ Paths						
▼ 1						
1	(:Link)-[:CONNECTS]-(:NE)	STEP	1378/1378/0/2756/0	0	2019/01/17 14:48:16	
2	240-load-topo-links-nodes	TRANS	0/0/0/0/0	0	2019/01/17 14:48:16	08.617"
3	240-load-topo-links-nodes	JOBENTRY	0/0/0/0/0	0	2019/01/17 14:48:16	
4	200-load-nodes-links	JOB	0/0/0/0/0	0	2019/01/17 14:48:13	12.128"
5	200-load-nodes-links	JOBENTRY	0/0/0/0/0	0	2019/01/17 14:48:13	
6	000-main-job	JOB	0/0/0/0/0	0	2019/01/17 14:48:12	19.323"

New stuff - www.kettle.be

- My blog
- Kettle Neo4j Remix downloads
 - Also with Beam support
- Documentation
- Links to key plugins
-

New stuff - Kettle Neo4j Logging

The screenshot shows the Spoon - Main job window. The job flow consists of several steps: START, Configure this, Setup SlaveServer nodes, Populate file queue, Repeat: Empty queue, and Repeat: Wait until finished. The 'Setup SlaveServer nodes' step is highlighted with a red arrow. Below the job flow, the 'Execution Results' pane shows a list of error messages, including 'Unexpected error' and 'JavaScript error: Hello, World!!! Something is going wrong here! (script#3)'. A blue arrow points from this pane towards the right-hand screenshot.

The screenshot shows the Execution history window for the job 'Main job'. It displays a table of error analyses with columns for #, Name, Type, R/W/I/O/R, and err. The table shows three main job entries and several sub-steps, with the 'Repeat: Empty queue' step highlighted in red. Below the table, the 'Name: Main job' section shows job details like Type: JOB, Date: 2019/11/22T13:22:06, and ID: 376a6250-a36d-4833-945c-885a327857fe. The 'Execution info cypher' section shows a Cypher query that filters for errors and returns the shortest path including metadata. The query is:

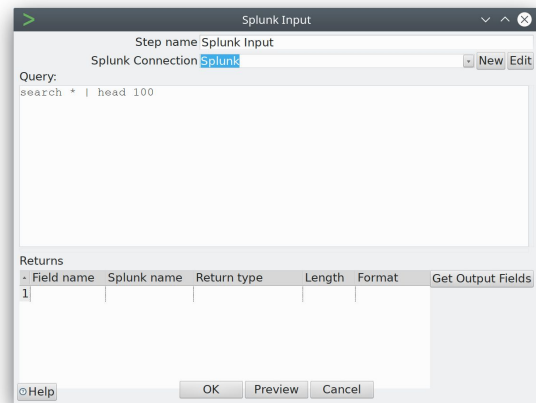
```
MATCH(terr:Execution { name: "Main job", type: "JOB", id: "376a6250-a36d-4833-945c-885a327857fe" })
WHERE terr.registrationDate IS NOT NULL
AND terr.errors > 0
AND size(terr)-[:EXECUTES]->(j)-0
RETURN p
ORDER BY size(RELATIONSHIPS(p)) DESC
LIMIT 5
```

New stuff - Kettle Splunk

- Read data from Splunk server
- No longer just an EE feature
- See my [Medium article](#)

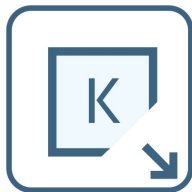


Splunk Input



New stuff - Kettle Kafka

- Fixing out of memory issue with PDI streaming steps
- Reverts back to basic Kafka Java API functionality
- Allows direct mapping for Avro messages
- Future improvements platform for type mappings



Kettle Kafka consumer

New stuff - Kettle Beam

- Integration of Apache Beam with Kettle
- see: <http://beam.kettle.be>
- Run your Kettle transformations on
 - Apache Spark
 - Apache Flink
 - Google Cloud DataFlow
 - Local runners: including Direct, Flink and Spark
 - Batch and Streaming



Kettle Conversations

- Analyzing state of Kettle, lack of innovation, frustrations
- On kettle-community slack channel
- With Apache Beam folks
- With the Apache Software Foundation
- With the Google folks
- With the Amazon Alexa folks (shout out Yuri and Phil)
- With the wonderful Neo4j people
- With several people at Hitachi Vantara
- ...



The Present





Project Hop

- www.project-hop.org
- a.k.a Apache Kettle, the desire to collaborate in a better way
- Shout-out to Hans who's wrapping up the day with it
- Shout-out to Jens who's going to explain how we can already collaborate on Kettle right now
- Contact me/us if you want to help out



Project Hop Goals

- Start Apache incubation as soon as possible
- Experiment with alternate ways of doing things
- Allow instability, rename, rebrand, sever ties, ...
- Improve documentation
- Tighten the architecture
- Lighten the load (down to 120MB already)
- Clean up APIs
- Merge projects (Beam, Environments, Unit testing, ...)
- Integration testing
- ...



The future



Lean

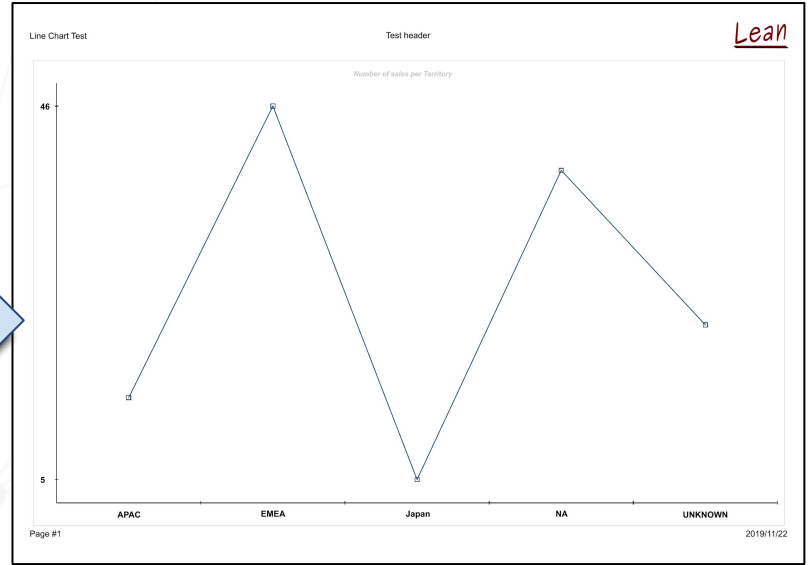
- Lean Enterprise Analytics
- Metadata driven presentation layer
 - Dashboards
 - Static and dynamic reporting
 - Storyboarding
 - ...
- Flexibility
- Auto-generated presentations
- Lightweight browser client
- Easy to use front-end

Lean

- Simple metadata structures
- Server side rendering in SVG

```
{  
  "name": "Line Chart Test",  
  "description": "Testing line chart aggregation",  
  "pages": [  
    {  
      "pageNumber": 1,  
      "width": 1123,  
      "height": 794,  
      "leftMargin": 25,  
      "rightMargin": 25,  
      "topMargin": 25,  
      "bottomMargin": 25,  
      "components": [  
        {  
          "name": "LineChart",  

```



Lean

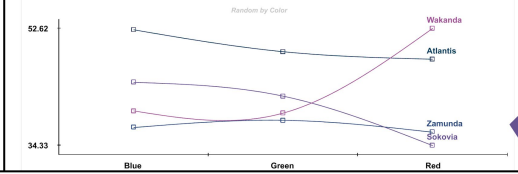
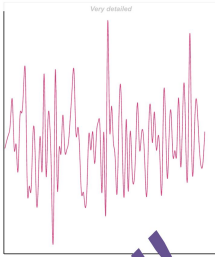
- Dynamic layouting
- Crosstabs
- Grouping
- Composite components
- Kettle plugin
- Neo4j plugin
- Node.js plugin
- ...

Combo (3000)

Layout test with charts basing position of crosstab

Lean

	Blue		Green		Red		Total
	N	Y	N	Y	N	Y	
Customer	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	
Adam Zapel	4.447	7 5.775	11 4.825	8 6.768	13 5.131	9 4.486	18 31.352 58
Alli Gaither	4.626	10 6.153	10 3.321	6 7.369	14 5.877	12 7.914	13 35.260 65
Anna Conda	2.547	6 5.001	9 2.614	5 4.747	13 2.437	5 1.722	4 19.147 42
Anne Teak	3.154	7 8.841	13 4.957	9 3.247	6 5.933	8 5.436	12 26.568 95
Barb Dwyer	5.689	11 8.223	15 2.636	6 2.738	6 0.398	2 4.344	8 24.820 48
Ben Dover	4.198	10 4.346	8 6.047	9 4.715	9 4.443	10 9.933	15 29.675 61
Bonnie Ann Clyde	7.361	15 5.451	9 8.128	12 4.349	9 5.213	12 2.272	7 32.766 64
Bug Light	2.334	5 5.168	11 1.607	3 4.369	12 2.717	5 5.249	9 21.445 45
Candace Spencer	3.728	9 5.179	11 5.123	9 2.348	4 5.743	12 5.152	9 27.274 54
Chris P. Bacon	1.397	4 6.450	11 18.452	17 4.064	8 5.612	9 4.017	7 32.792 56
Dixie Normous	2.258	5 3.968	6 5.738	11 2.387	5 7.429	13 5.845	9 27.618 49
Doug Hole	2.397	5 2.782	9 5.164	13 4.369	10 3.558	8 6.648	10 24.918 55
Earl Lee Riser	4.484	9 3.460	6 5.019	10 2.276	8 5.198	12 5.431	9 25.819 54
Jed I Knight	6.396	10 4.434	9 4.724	10 5.039	13 6.284	11 3.934	6 38.732 59
Justin Slicer	2.392	6 2.993	6 6.088	15 5.343	8 5.785	11 6.899	12 29.520 58
Ken Hurt	4.840	8 6.879	14 4.343	8 3.698	9 4.264	9 5.552	10 28.776 58
Kent C. Strait	7.544	13 9.992	19 4.037	8 4.368	8 3.893	5 3.826	9 33.652 62
Mike Liboris	6.551	12 3.589	7 6.224	12 3.748	7 4.592	10 5.253	6 38.056 57
Total	75.527	152 97.965	184 92.050	171 76.667	162 81.946	163 89.234	168 513.389 1800



LEAN!

2019/11/22

Group composite (8000)

A group repeating a composite with a label and a crosstab, data filtering

Lean

Country: Atlantis

	Blue		Green		Red		Total
	N	Y	N	Y	N	Y	
Customer	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	
Adam Zapel	7.708	14 7.724	15 7.292	14 6.827	15 5.346	10 6.527	14 41.477 82
Alli Gaither	8.219	17 6.069	14 8.477	14 5.666	14 7.218	15 7.848	12 42.681 86
Anna Conda	4.827	18 6.688	14 3.570	7 6.339	15 5.563	10 3.796	18 29.784 66
Anne Teak	9.891	19 6.757	13 6.924	13 9.682	16 5.872	13 7.755	16 45.881 90
Barb Dwyer	6.346	15 8.332	18 6.240	13 4.543	9 2.483	6 6.223	11 34.187 72
Ben Dover	6.179	14 4.913	13 4.459	13 6.758	12 7.887	15 6.174	20 39.283 87
Bonnie Ann Clyde	10.488	19 7.819	13 8.958	15 5.327	11 7.539	19 4.881	13 44.925 90
Bug Light	6.156	12 7.231	14 4.774	11 6.820	17 4.554	10 7.158	12 35.893 76
Candace Spencer	5.178	18 7.195	16 4.982	9 6.086	13 10.614	15 8.704	16 43.688 83
Chris P. Bacon	2.828	5 6.172	18 10.426	18 18.618	18 7.224	11 5.812	9 42.280 71
Dixie Normous	4.674	18 6.986	12 7.813	15 7.910	14 8.824	15 10.366	16 45.772 82
Doug Hole	5.371	13 6.918	16 18.050	21 4.913	12 6.184	11 9.877	16 42.521 89
Earl Lee Riser	5.411	13 6.445	13 6.356	11 6.864	18 9.786	20 5.469	18 40.251 85
Jed I Knight	9.371	18 3.898	9 6.196	14 3.087	13 8.397	16 1.975	5 33.445 75
Justin Slicer	4.858	13 7.883	6 5.537	10 7.079	14 8.076	10 8.062	19 41.458 80
Ken Hurt	5.246	11 9.118	19 5.948	12 5.715	14 6.297	11 10.188	18 42.511 85
Kent C. Strait	10.212	18 9.883	28 18.855	20 5.468	13 8.894	15 4.896	18 49.318 96
Mike Liboris	8.492	15 7.283	13 6.913	13 5.768	12 4.817	10 9.848	16 42.231 79
Total	119.020	246 122.197	248 138.787	252 116.401	250 124.607	244 124.462	243 737.396 1483

Country: Sokovia

	Blue		Green		Red		Total
	N	Y	N	Y	N	Y	
Customer	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	Sum Count	
Adam Zapel	8.598	14 9.095	16 7.265	14 9.287	18 8.449	13 6.368	13 47.749 86
Alli Gaither	7.519	19 9.913	19 16.162	19 7.161	6 10.821	16 9.350	18 48.379 94
Anna Conda	5.856	12 7.438	15 5.835	12 6.984	16 6.588	12 4.159	18 36.781 77
Anne Teak	7.438	16 9.186	15 3.924	9 4.497	12 5.272	13 9.108	16 39.416 81
Barb Dwyer	6.492	15 9.374	18 6.619	13 6.843	12 2.397	6 7.491	13 37.746 77

Page #1

2019/11/22

Lean

- Will be released as APL 2.0
- When we're ready :-)

Lean Solutions

- A company, foundation style
- Bart, Hans, Matt investing
- ... if not always participating
- Aimed at supporting the Hop, Lean & Neo4j combo to anyone who needs it
- "Soon"

A network diagram with a central node at the bottom, branching out into a fan shape. The nodes are represented by circles of varying sizes, and they are connected by thin lines. The background is a gradient from blue on the left to yellow on the right.

THANK YOU!!





Questions?

