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

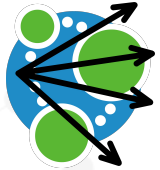
- One of the coolest things ever in ML
- Welcome Mark Hall and Ken Wood
- Presentation after the coffee!

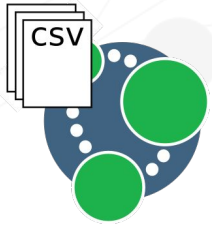
Improvements

- 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

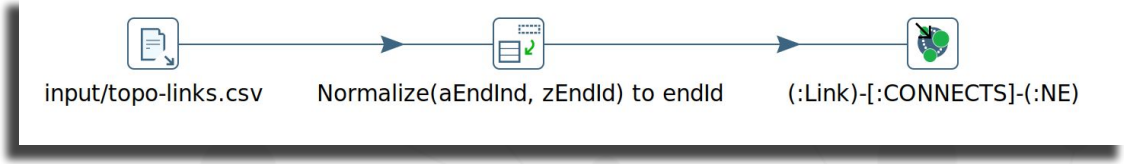


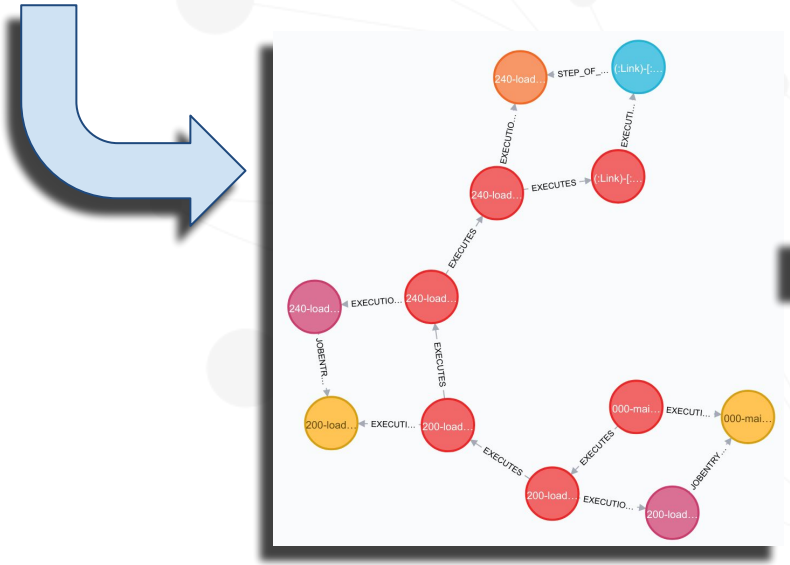


Kettle Neo4j Logging

- 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





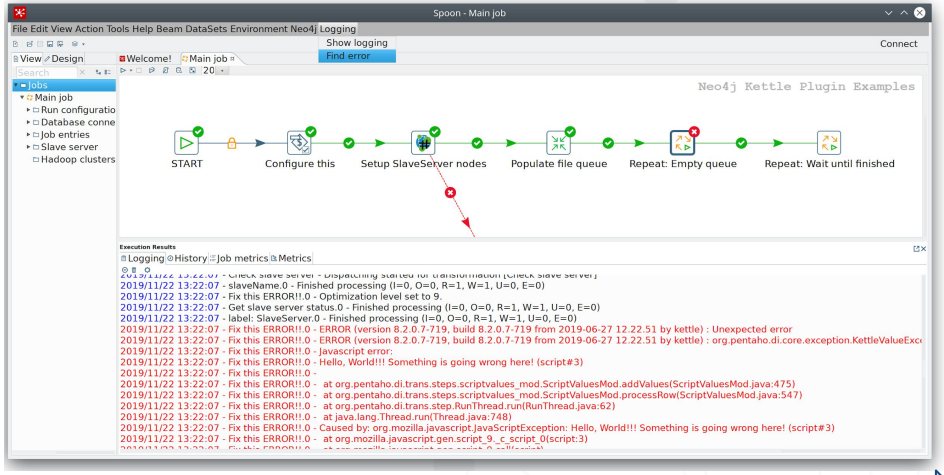
Topic: Execution history of step '(:Link)-[:CONNECTS]-(:NE)'
Parent: 240-load-topo-links-nodes (TRANS)

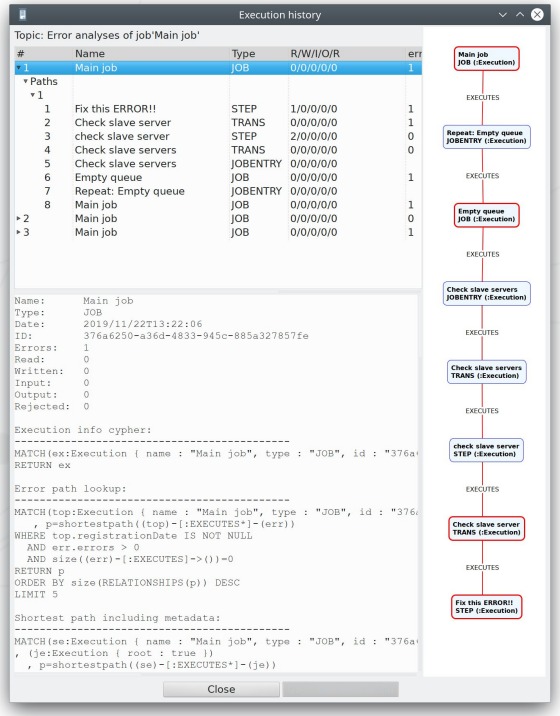
#	Name	Type	R/W/I/O/R	errors	date	duration
1	(:Link)-[:CONNECTS]-(:NE)	STEP	1378/1378/0/2756/0	0	2019/01/17 14:48:16	
Paths						
1						
1	(:Link)-[:CONNECTS]-(:NE)	STEP	1378/1378/0/2756/0	0	2019/01/17 14:48:16	
2	240-load-topo-links-nodes	TRANS	0/0/0/0/0	0	2019/01/17 14:48:16	08.617"
3	240-load-topo-links-nodes	JOBENTRY	0/0/0/0/0		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		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
-

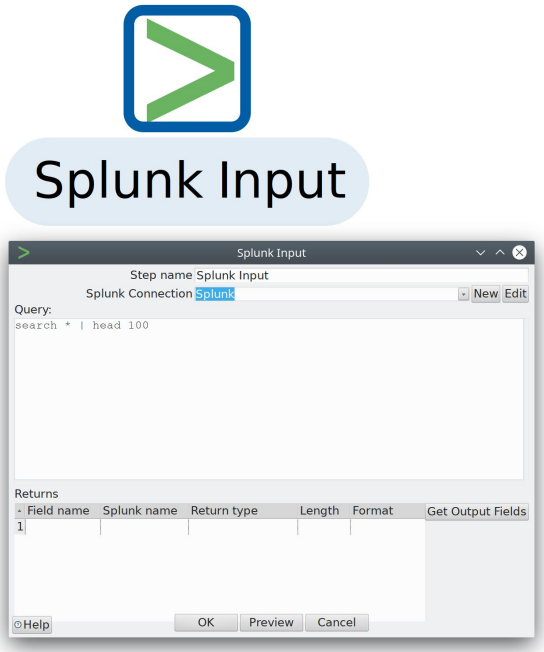
New stuff - Kettle Neo4j Logging





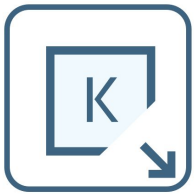
New stuff - Kettle Splunk

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



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

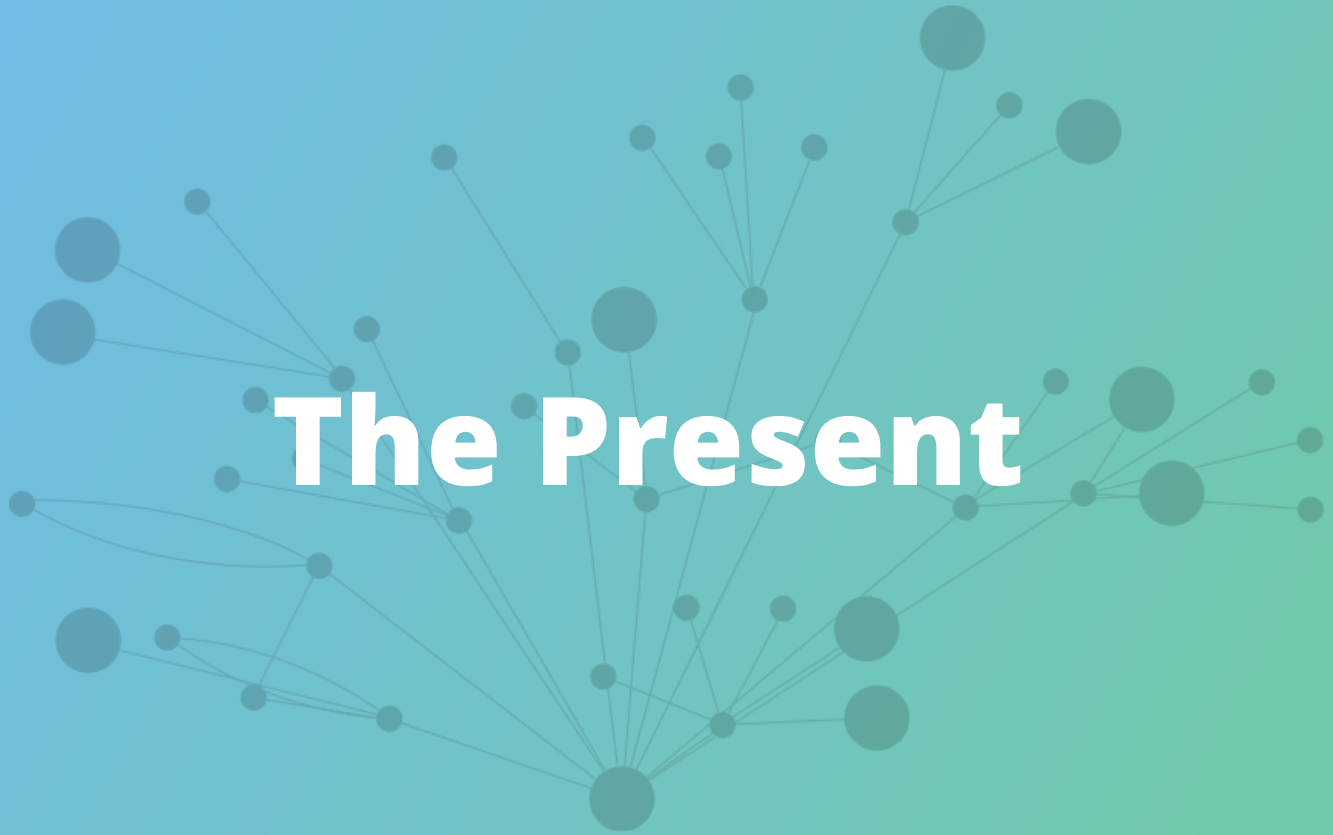
New stuff - Kettle Beam

- 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

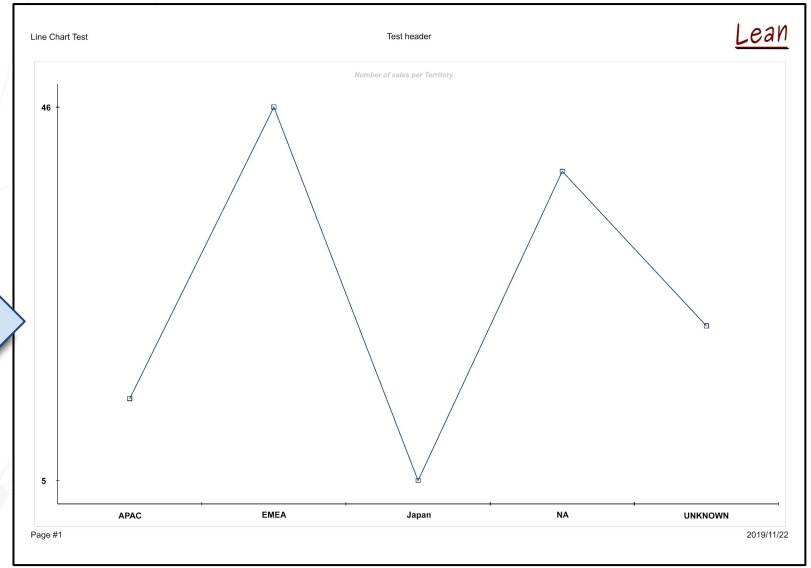
Lean

- 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

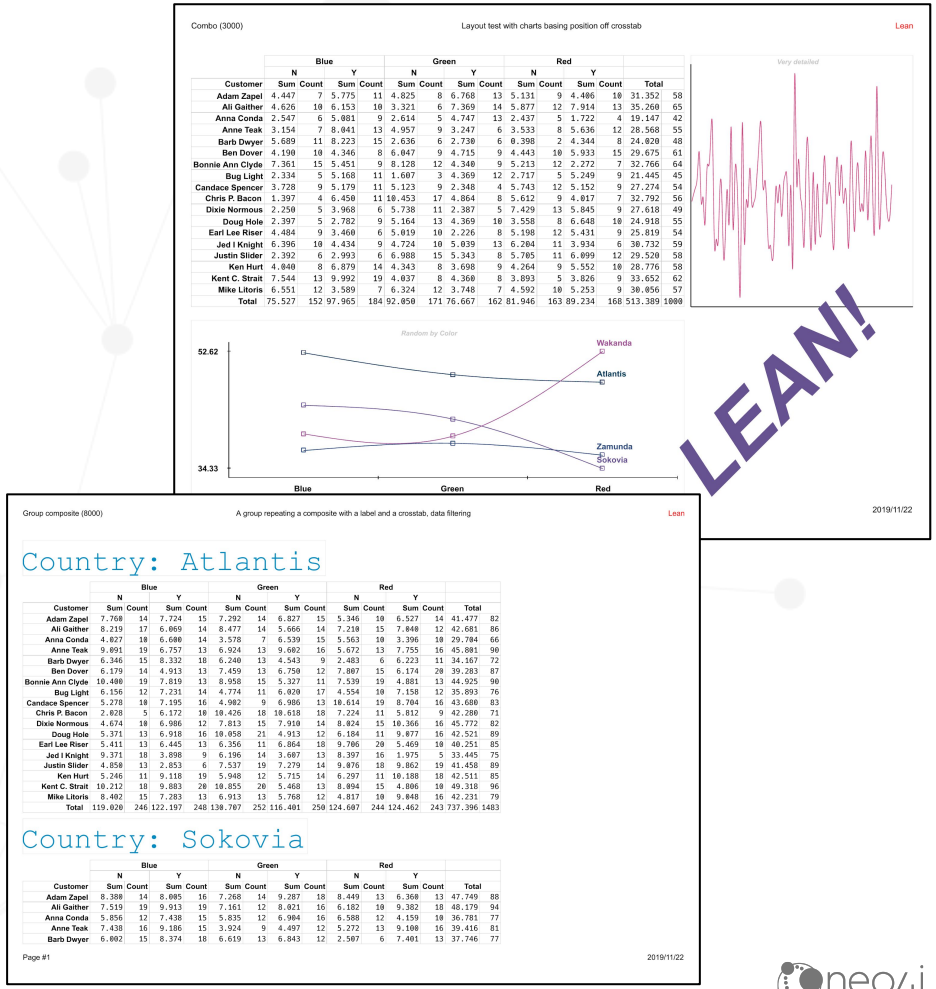
- Simple metadata structures
- Server side rendering in SVG

```
{
  "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
- …



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”

THANK YOU!!

Questions?