



## Pentaho Data Integration version 2.4.0

Integrating more, easier & faster!

*Key differences with Kettle 2.3.0*

*List of changes on 20-01-'07*

Compiled by Matt Casters, [mcasters \(at\) pentaho.org](mailto:mcasters@pentaho.org)

Send additional changes you found to this address.

## Index

1. Changes summary.....	3
1.1. Preface.....	3
1.2. Overview.....	4
2. General changes.....	5
2.1. Spoon enhanced, RIP Chef.....	5
2.2. A new welcome screen.....	5
2.3. Variable support with visual feedback.....	5
2.4. A new execution configuration dialog.....	6
2.5. Slave servers & clustering.....	6
2.6. Databases.....	7
2.7. Shared objects.....	9
2.8. Repository improvements.....	10
3. Spoon.....	11
3.1. Extra options.....	11
3.2. Search meta-data.....	12
4. Steps.....	13
4.1. Text file input.....	13
4.2. Get System Info.....	13
4.3. Generate rows.....	14
4.4. Text file output.....	14
5. Source code improvements.....	15
5.1. A few extra lines of code.....	15
5.2. Trackers.....	15
5.3. Commiters.....	15
5.4. Other contributors.....	15

# **1. Changes summary**

## **1.1. Preface**

It's been almost 7 months since the last major revision of the Pentaho Data Integration software (a.k.a. Kettle). Again, we have had a lot of feedback, massive amounts of downloads and in general it's fair to say that the community is thriving on the new forum to which we migrated.

## 1.2. Overview

These are the most notable changes that have been made:

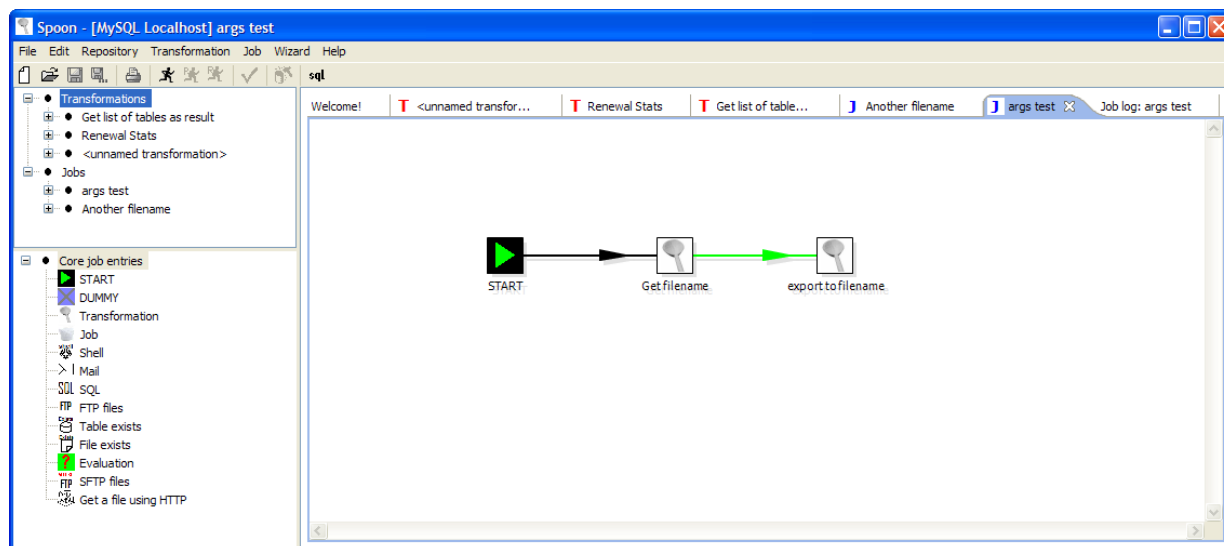
- Improved Performance and Scalability
  - Remote monitoring and execution of Jobs and Transformations
  - Clustering support (MPP)
  - Database partitioning support (not to be confused with table partitioning)
  - Numerous memory and performance improvements
- Enhanced database support
  - Improved quoting algorithms
  - separation of schema/owner and table names throughout the steps
  - Databases connection pooling support
  - Clustering support
  - 25 database types (+generic) are now supported. The new database types that are supported now are: *Teradata*, *Oracle RDB*, *H2*, *Netezza*, *IBM UniVerse* and *SQLite*.
- Ease of Use
  - Spoon is now an integrated designer for Jobs and Transformations
  - Design and edit multiple transformations simultaneously
  - Repository improvements including the ability to sort contents by name, user, object type and modified date
  - Variable enhancements including added support in many step types and visual indicators on all fields that support variables
- Powerful New Steps
  - High performance, expression-based Javascript step
  - Add XML step for building sophisticated XML constructs from stream fields
  - Merge Join step for advanced joins like INNER, LEFT OUTER, RIGHT OUTER, AND FULL OUTER
  - Fast Sorted Merge join for merging multiple streams sorted on the same key
  - HTTP Client lookup to dynamically retrieve parameters using a web service
- Miscellaneous
  - Remote monitoring of transformation and job execution
  - Ability to share database connections using XML without having to use the repository
  - Ability to roll-back entire transformations on error (unique connections)
- By the numbers
  - >200.000 downloads of 2.3.0 and 2.3.1 (mostly 2.3.1)
  - 4000 SVN commits
  - >1000 forum threads
  - 262 closed bugs with an average closing time of 29 days.
  - 51 feature requests implemented with an average implementation time of 100 days.
  - 10 developers committing to SVN

## 2. General changes

### 2.1. Spoon enhanced, RIP Chef

As you can see from the image below you can now load multiple transformations and jobs into a single Spoon editor. This removed the need for the Chef job editor and makes it easier to edit larger jobs.

Not only can you load jobs and transformations, but you can also run multiple jobs & transformations, monitor remote slave servers and view execution history in the same environment.



### 2.2. A new welcome screen

TODO: Make screenshot

### 2.3. Variable support with visual feedback

Username  \$

## 2.4. A new execution configuration dialog

This dialog combines local, remote and clustered execution as well as preview, log settings, safe mode, arguments and variable support in one dialog.

**Execute a transformation**

Local, remote or clustered execution

☒ Local execution ☐ Execute remotely ☐ Execute clustered

Remote host:

☒ Post transformation  
☒ Prepare execution  
☒ Start execution  
☐ Show transformations

☒ Preview

#	Step	Nr of rows to preview
1	tables.txt	5000
2	Copy rows to result	5000
3	get list of tables	5000
4	Select values	0

☒ Enable safe mode

Log level:

Replay date (yyyy/MM/dd HH:mm:ss):

Arguments

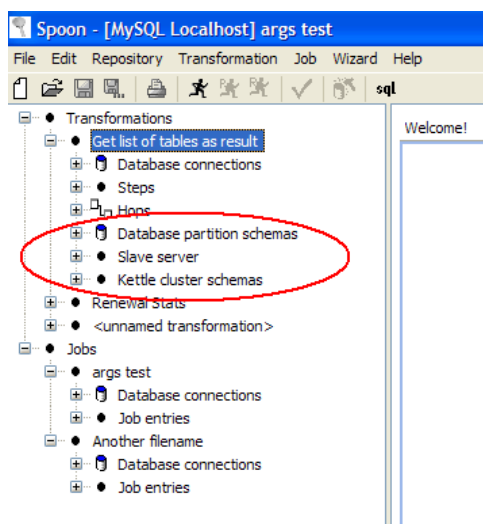
Argument	Value
1	

Variables

Variable	Value
1 FILENAME	C:\Temp\file.txt

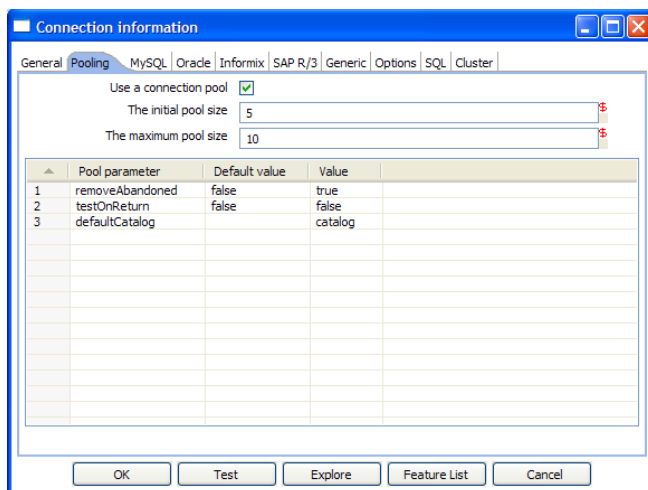
## 2.5. Slave servers & clustering

With this new release we move into the fascinating realm of massive parallel computing. It is now not only possible to execute a transformation on a remote server, but also to execute steps in a clustered way. By clustered we mean in parallel on more than one server. To allow this to work we created a small web-server called "Carte" that will accept input from either Spoon (remote & clustered execution) or from the Transformation job entry. (clustered execution)

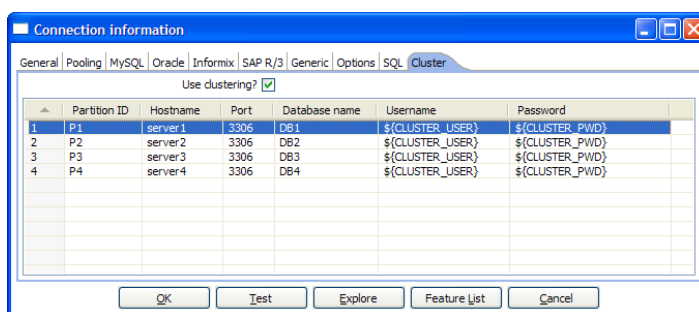


## 2.6. Databases

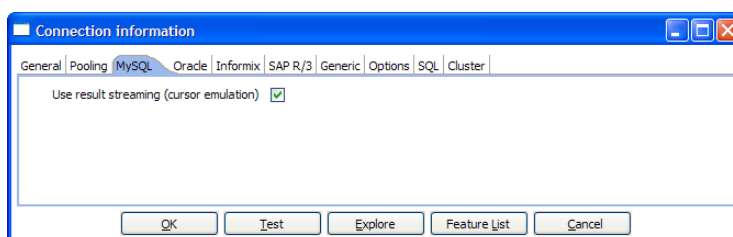
- Teradata support (<http://www.teradata.com/>)
- Oracle RDB support (<http://www.oracle.com>)
- H2 support (<http://www.h2database.com>), requested first by Kent Andrews
- Netezza support (<http://www.netezza.com>), dialect and driver by Biswapesh Chattopadhyay for Google
- IBM UniVerse support (<http://www-306.ibm.com/software/data/u2/universe/>), testing & driver by Frans van Dortmont)
- SQLite support (<http://www.sqlite.org/>), implemented with help from Tom Gleeson  
<http://gobansaor.wordpress.com/2006/12/18/sqlite-jdbc-and-kettle-pentaho-data-integration-etl/>
- Connection pooling support added, written by 陈萍 (a.k.a. Apple : <[chenping@gxlu.com.cn](mailto:chenping@gxlu.com.cn)>), feature request 3838 by Jeremy Haile <[javaforge@jhaile.fastmail.fm](mailto:javaforge@jhaile.fastmail.fm)> was also implemented.



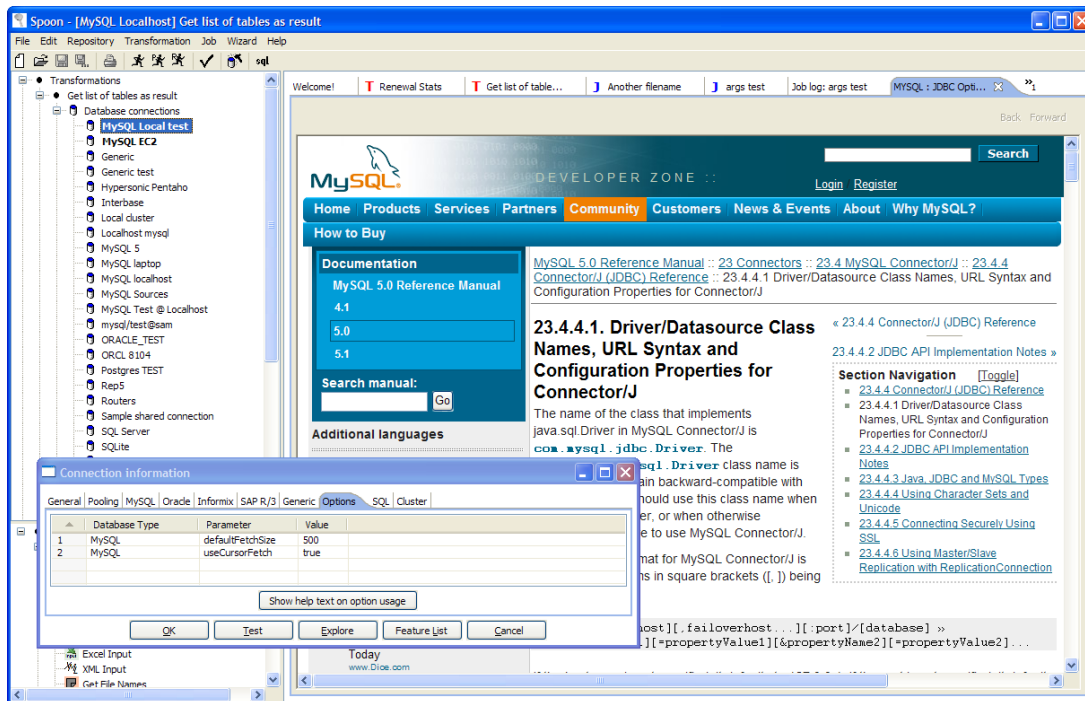
- Clustering/partitioning support to allow data to be split up among databases



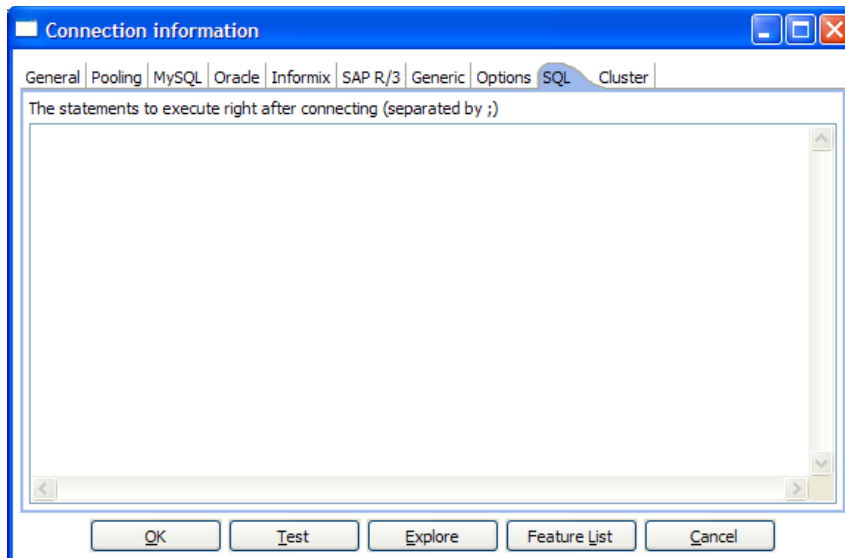
- MySQL: Allow cursor emulation (a.k.a. result streaming) to be turned off



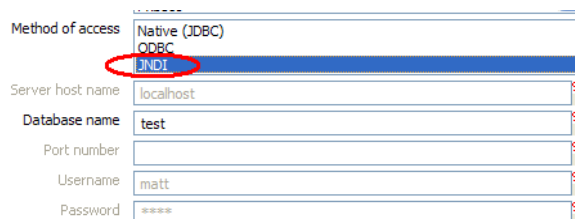
- Clicking the options help button now opens a new browser tab in Spoon:



- Right after connecting, it is now possible to run a number of SQL commands. This is sometimes needed for various reasons such as licensing, configuration, logging, tracing, etc.



- JNDI connections are now also supported (code by James Dixon, Pentaho)



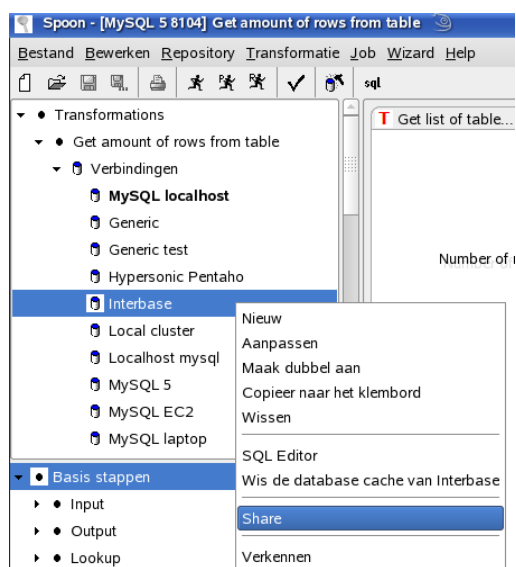


## 2.7. Shared objects

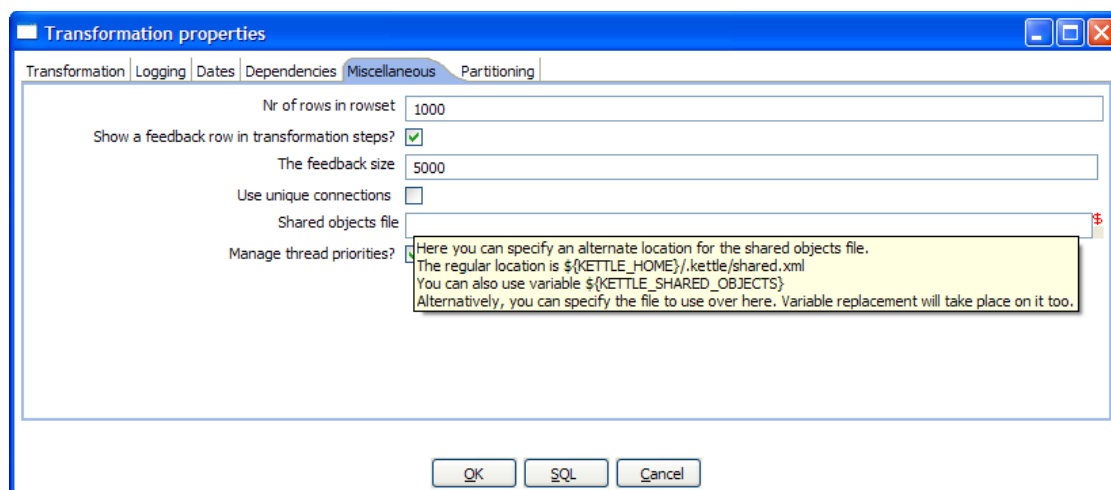
Various objects can now be placed in a shared objects file. This file (default: \$HOME/.kettle/shared.xml) can be used to define:

- Database connections
- Steps
- Slave servers
- Partition schemas
- Cluster schemas

On these objects you can simply click right and select “Share”.

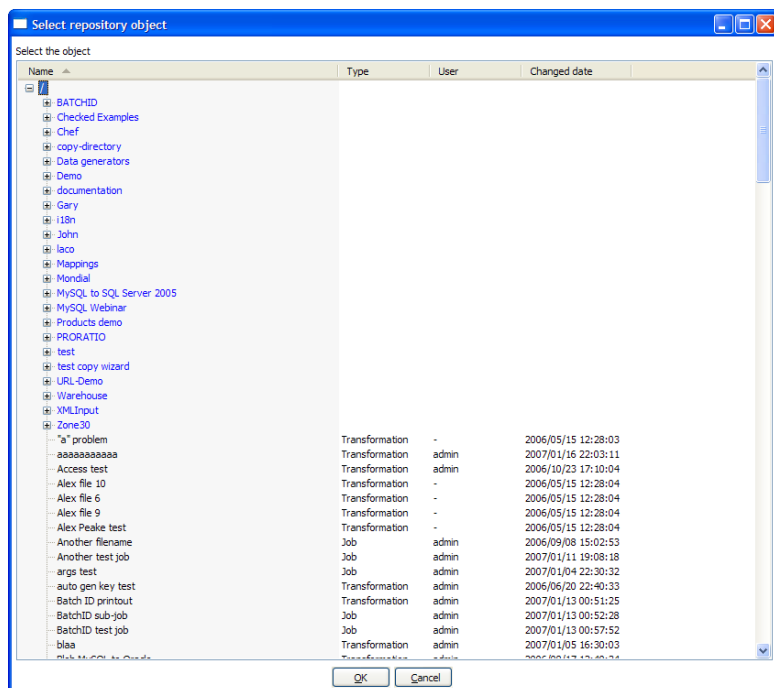


Doing so will place the object in the shared objects file. The location of the file is configurable:

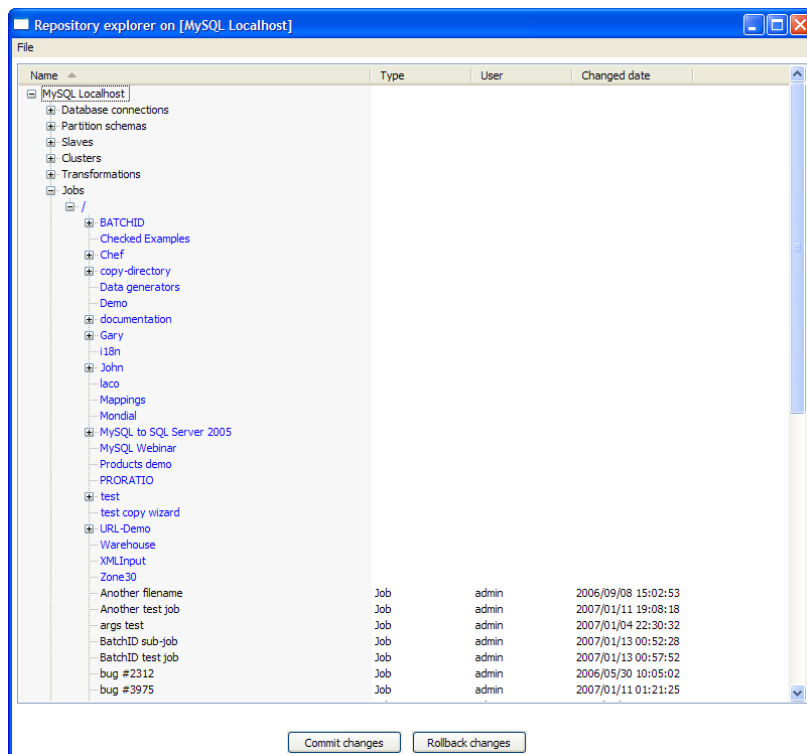


## 2.8. Repository improvements

We included extra information on the jobs and transformations stored in the repository in as well the repository explorer as the File/Open operations. This makes it easier for you to search for a particular transformation or job:



The repository also contains support for this as well as for the new objects like slave servers:

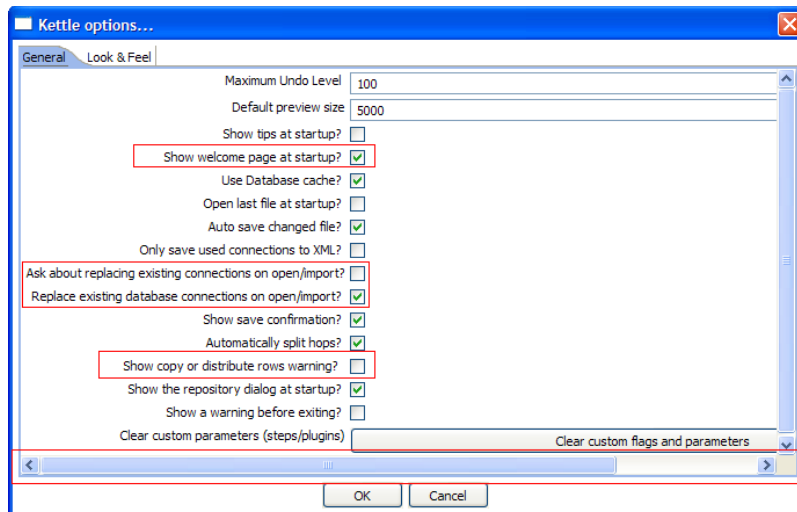


## 3. Spoon

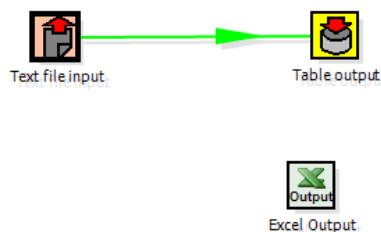
For a list of the changes that were done in the steps & job entries, please see the corresponding chapters below.

### 3.1. Extra options

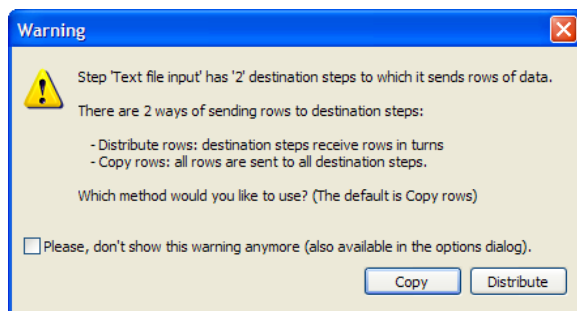
These are the new Spoon interface options that were added:



Most of the new options explain themselves. However, here are the details on the “copy/distribute rows warning” option. If you have a transformation like this:

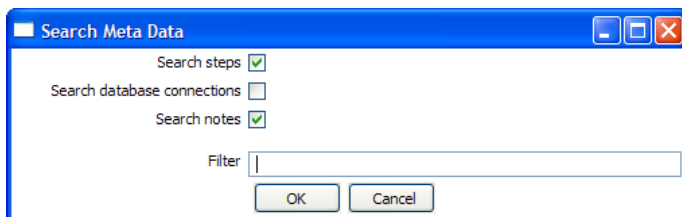


At the time you link step “Text file input” to “Excel Output”, you will be shown this dialog:



Apart from this, we made the dialog content scroll if it is sized too small. We do this to prevent options from being hidden involuntarily.

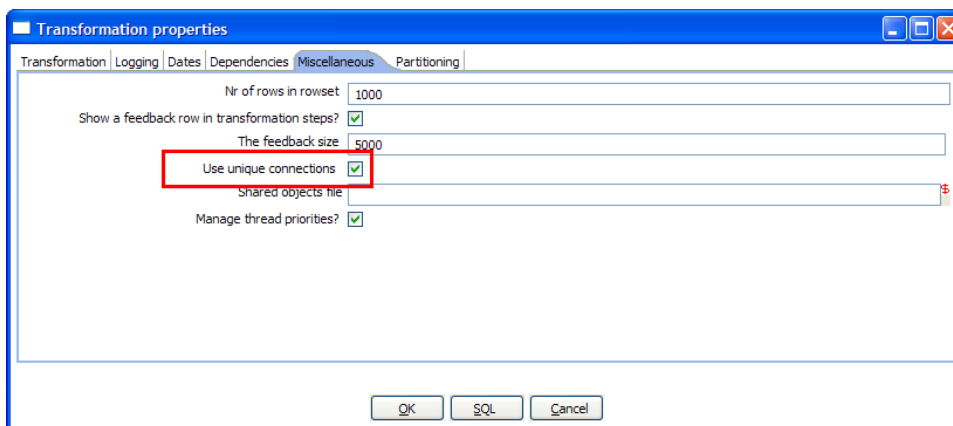
## 3.2. Search meta-data



This option was introduced in 2.3.0 but will now search through all transformations and jobs that are loaded in Spoon.

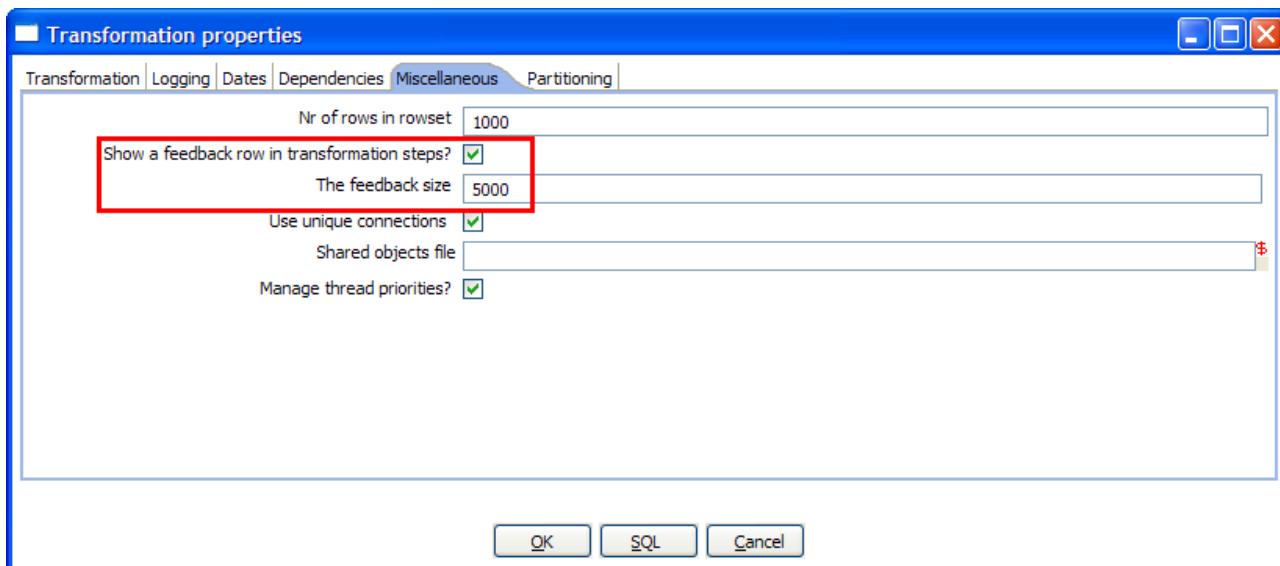
## 3.3. Unique connections / transactional fitness

We have made a serious attempt to make sure that failing transformations can be rolled back completely if this is desired. Typically you might use this for smaller data volumes and/or data migration/integration exercises. This feature is accomplished by opening one unique connection per defined and used database connection in the transformation.



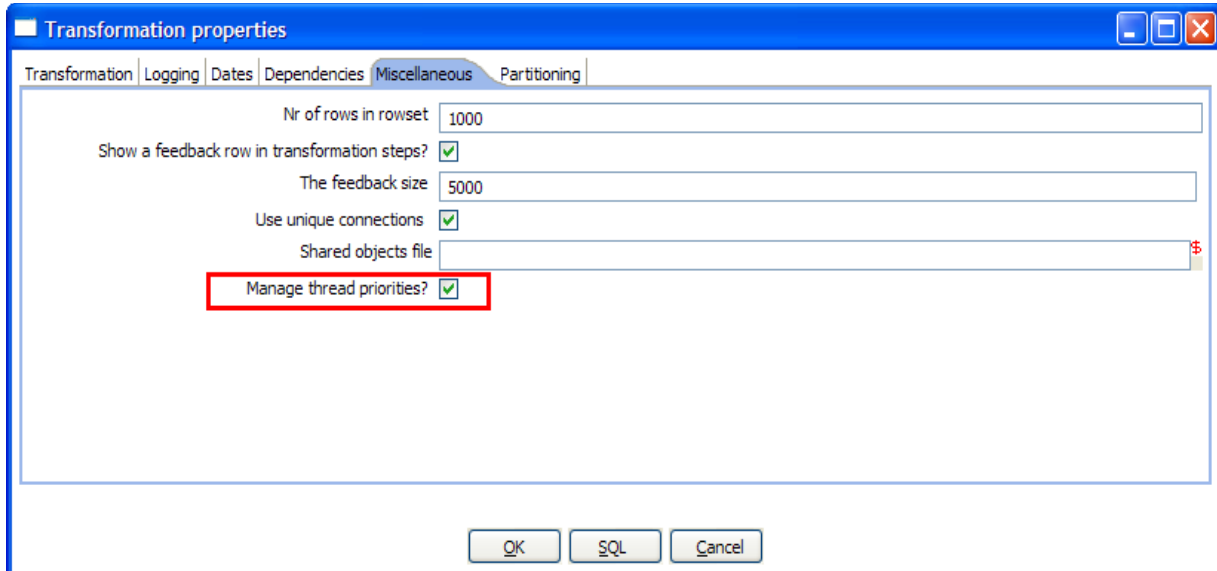
## 3.4. Feedback configurations

While processing millions of rows it logging can grow to large volume with a feedback every 5000 rows. This is now configurable in the transformations settings. You can turn it of or change the interval or feedback size.



### 3.5. Thread priority management

Since a few years back, Kettle transformations change Java thread priorities based on the number of input and output rows in the respective “rowset” buffers. However, in certain simplistic situations, this is costing more time than it's worth and therefor you can now turn it off if you want.



## 4. Modified steps

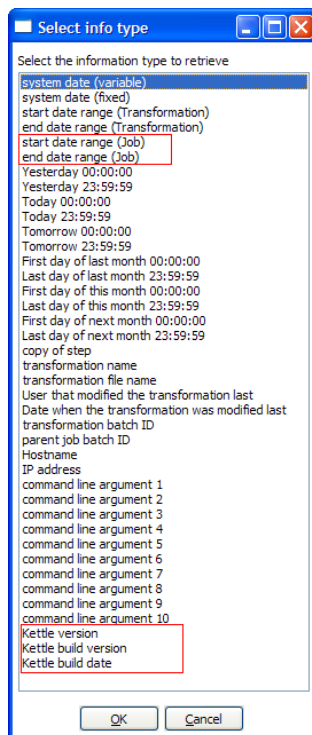
### 4.1. Text file input

Paged layout (printout)? ☐ Number lines per page 80  
 Document header lines 0  
 Compression None  
 No empty rows ☒  
 Include filename in output? ☐ Filename fieldname  
 Rownum in output? ☒ Rownum fieldname rownr  
 Rownum by file? ☒  
 Format DOS

- Allow the method of compression to be specified as either Gzip or Zip.
- Allow the row number to be reset per file.

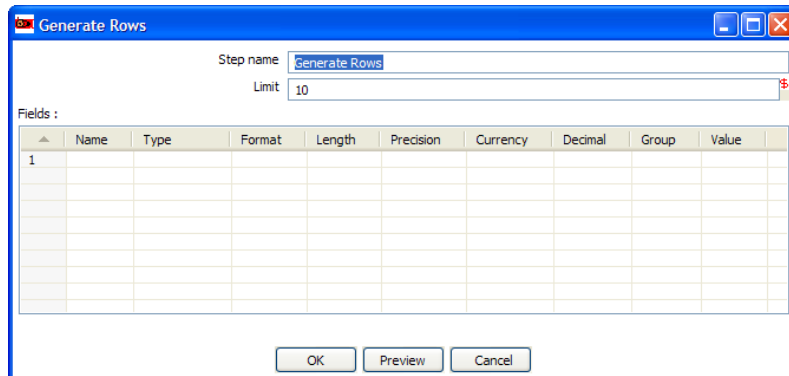
### 4.2. Get System Info

We gave access to the job start and end of the date range as well as the Kettle specifics like version, build date and build version.



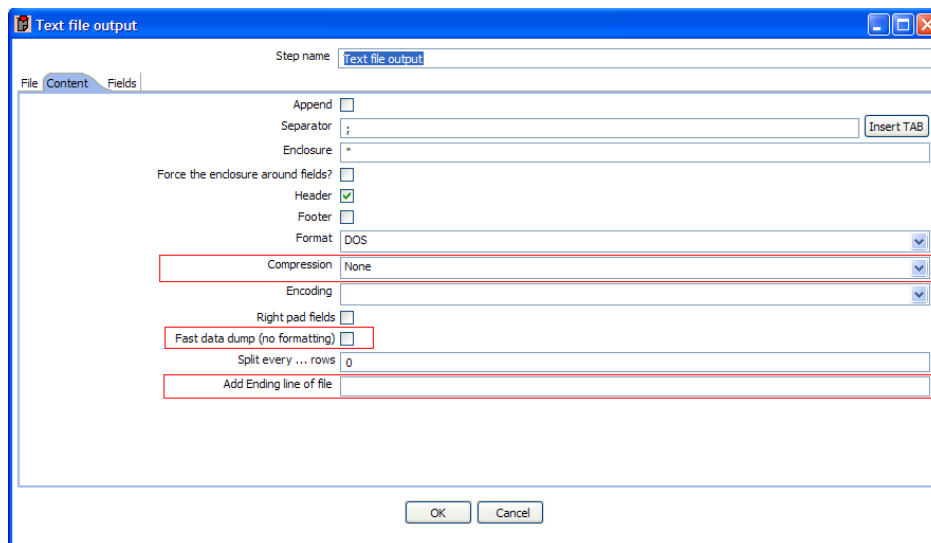
### 4.3. Generate rows

Now accepts for the limit to the number of generated rows a variable:



### 4.4. Text file output

We added support for Gzip and Zip compression as well as a “Fast data dump” option to speed up dumping large amounts of data to a text file. We also added the option to come up with an alternative ending for your text.



### 4.5. Table Output, Insert/Update, Update, Delete, Database Lookup, Dimension Lookup/Update, Combination lookup/update

Added the schema name to improve precision in the quoting and allow for table-names with dots '.' in it. (like on certain CRM systems like Axapta)

### 4.6. De-serialize from / Serialize to file

Re-branded Cube-input/output steps. (same as before but optimized for speed)

## 4.7. Stream Value Lookup

We have included the option to encode rows of data to preserve memory. These options, especially 'Preserve memory' and 'Key and Value are exactly one integer field' lead to enormous gains in memory efficiency. While these encodings can lead to a slight decrease in performance, it is in general still worth to do it. We have seen tens of millions of rows being loaded in less than a GB of RAM this way. This opens up possibilities that were not available previously.

Step name:

Source step:

The key(s) to look up the value(s):

	Field	LookupField
1		

Specify the fields to retrieve :

	Field	New name	Default	Type
1				

☒ Preserve memory (costs CPU)  
☐ Key and value are exactly one integer field  
☐ Use sorted list (i.s.o. hashtable)

The sorted list option is less performing than the others but will lead to the possibility to allow for other-than-equal key comparisons in the next version of Kettle.

## 4.8. Sort rows

The sorting algorithm has been optimized and will now perform a lot better than before, expect 100% or more speed gains.



## 4.9. Add sequence

The old behavior of the Add Sequence step was that if you had multiple steps in a single transformation that generated the same value name, the sequence would give unique numbers to these steps. That is because it uses the same counter. Now you can set the name of the counter for if you do not want that behavior.

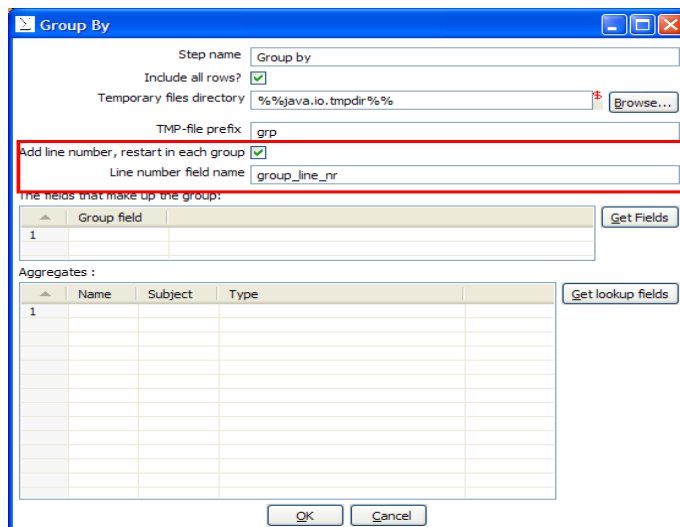
The screenshot shows a dialog box titled "Get sequence value from database". It has two main sections: "Use a database to generate the sequence" and "Use an transformation counter to generate the sequence".

In the "Use a database to generate the sequence" section, the "Use DB to get sequence?" checkbox is unchecked. The "Connection" field has a dropdown menu and "Edit..." and "New..." buttons. The "Schema name" and "Sequence name" fields are empty.

In the "Use an transformation counter to generate the sequence" section, the "Use counter to calculate sequence?" checkbox is checked. The "Counter name (optional)" field is highlighted with a red rectangle. The "Start at value" field is set to 1, the "Increment by" field is set to 1, and the "Maximum value" field is set to 9999999. At the bottom are "OK" and "Cancel" buttons.

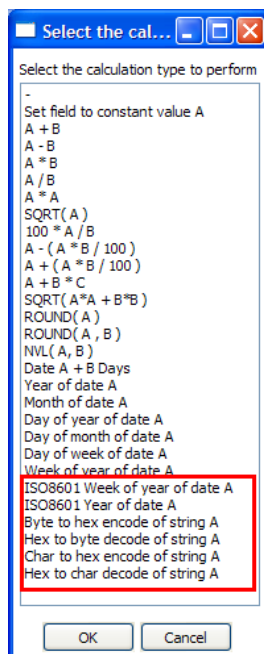
## 4.10. Group by

In case you allow all rows to pass and the aggregated information is added to all rows in the defined groups, you can now also add a line-number that restarts at 1 in each new group.



## 4.11. Calculator

The calculator can now perform a few extra calculations.



## 4.12. Denormaliser

It is now also possible to use variables in stead of hard-coded entries for the key-values.

Step name: Row denormaliser

The key field:

The fields that make up the grouping:

Group field
1

Get Fields

Target fields:

Target fieldname	Value fieldname	Key value	Type	Format
1				
2				
3				

Get lookup fields

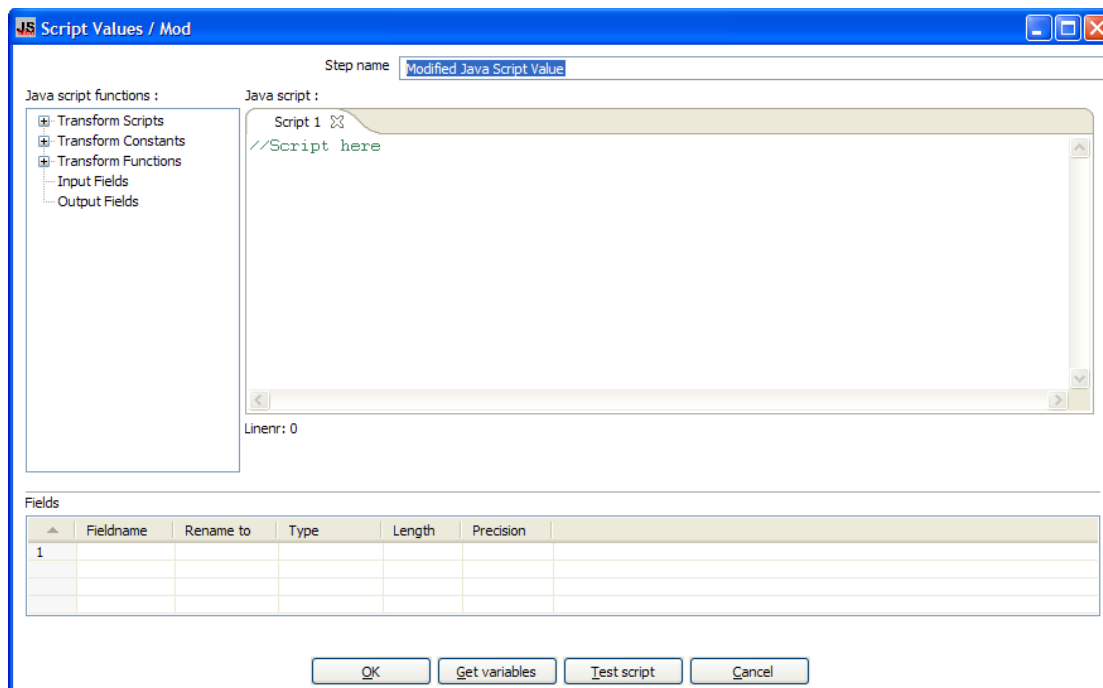
Enter CTRL-SPACE to select a variable to insert

OK Cancel

## 5. New steps

### 5.1. Modified Javascript Value

Martin Lange from company Proconis (<http://www.proconis.de/>) rewrote the existing “Javascript values” looking for better ease of use in the GUI and better performance. This has resulted in the “Modified Javascript Value” step.

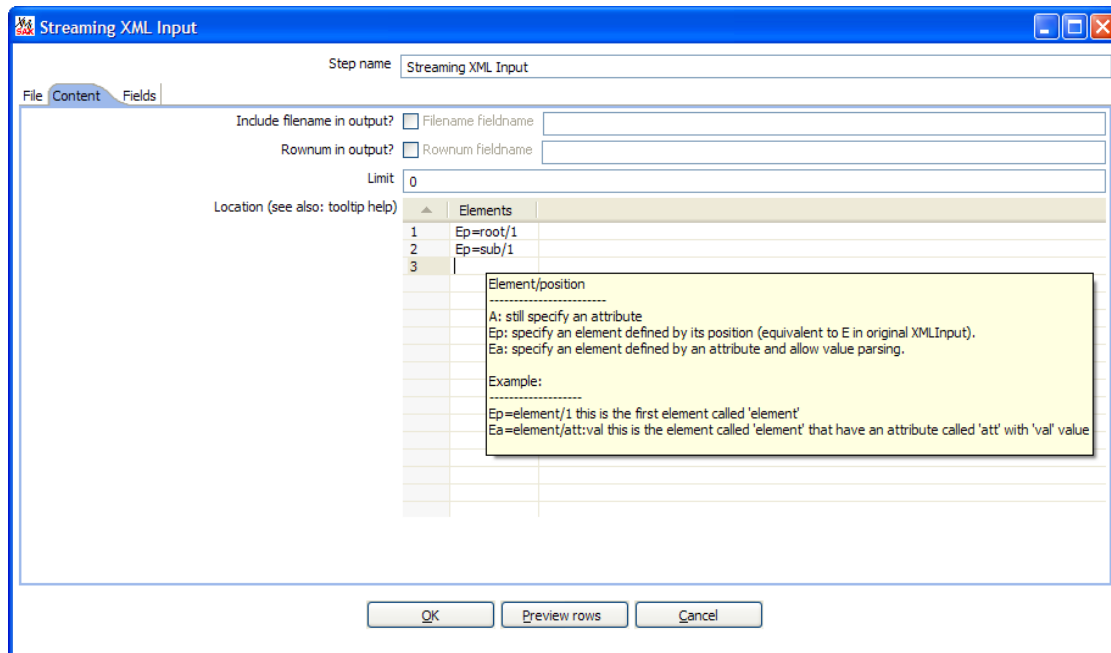


As you can see from the screenshot above it is now possible to have multiple separated script tables, but also “Start” and “End” scripts (before and after).

Color coding is also available for known keywords, functions, constants, comments, etc.

Because the executable code is not 100% compatible (for performance reasons) we have opted to keep the regular and the modified Javascript steps alongside each other.

## 5.2. Streaming XML Input



The purpose of this step is to provide value parsing.

This step is based on SAX parser to provide better performances with larger files.

It is very similar to XML Input, there are only differences in content and field tabs.

### **Content**

Location specifies the path by way of elements to the repeating part of the XML file. The new thing, is that in element column you specify the position of the element as described below.

### **Field**

There is a new grid that allows specifying an association between an element name and its defining attribute name. This makes the step able to retrieve the good fields when pressing "get fields" button.

The position column is also different it works like described above

### **Element/position**

A: still specify an attribute

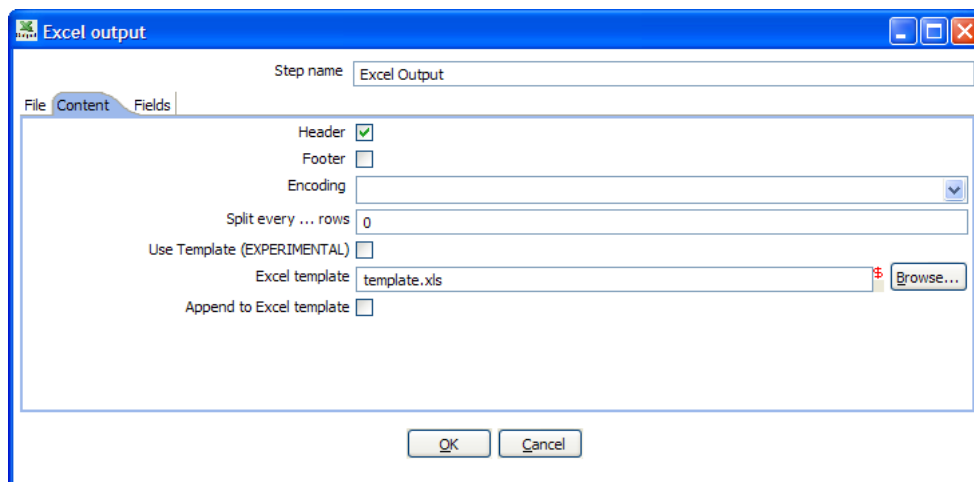
Ep: specify an element defined by its position (equivalent to E in original XML Input).

Ea: specify an element defined by an attribute and allow *value parsing*.

## 5.3. Excel Output

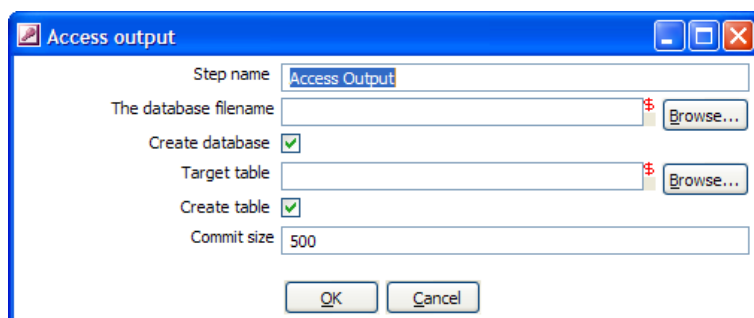
Kettle has no intention whatsoever to go and replace functionality found in complex reporting engines.

That being said, once in a while it is useful to be able to output to simple Excel sheets natively. Because of that the Excel Output step was written. It only allows for very basic formatting options, but keep in mind that that is not the intent of this step.



## 5.4. Access Output

This falls in the same category as the Excel output step: it can sometimes be handy to create new Access database files, but don't expect us to add complex functionality to this.

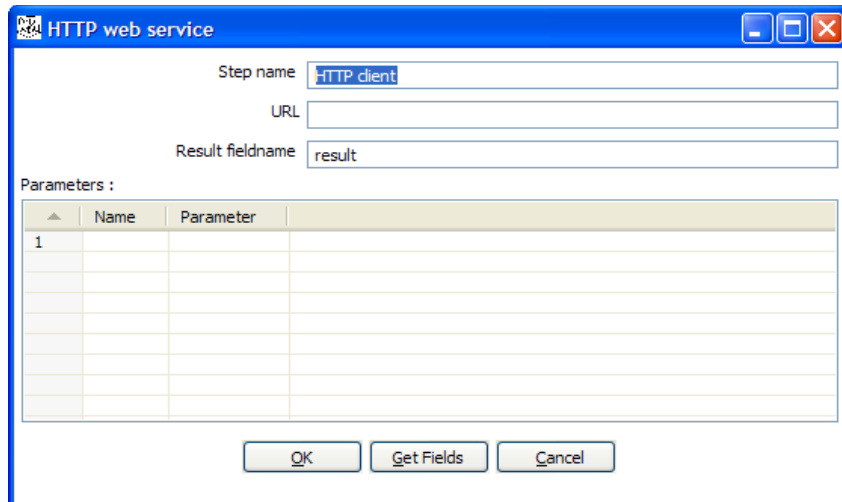


## 5.5. HTTP Web service

The HTTP Web service performs a very simple call to a base URL with options appended to it like this:

`http://<URL>?param1=value1&param2=value2&...`

The result is stored in a String field with the specified name.

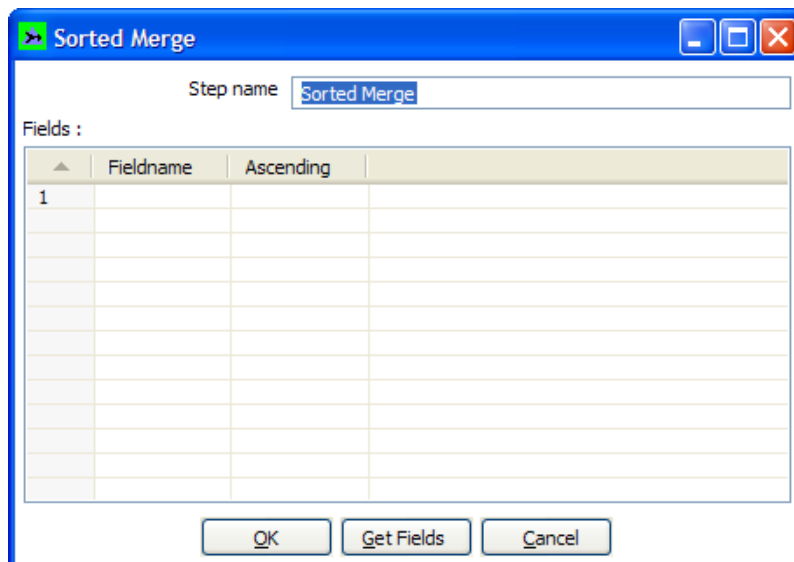


The screenshot shows the 'HTTP web service' dialog box. It has a title bar with a blue background and standard window controls. Inside, there are three text input fields: 'Step name' with the value 'HTTP client', 'URL' which is empty, and 'Result fieldname' with the value 'result'. Below these is a section labeled 'Parameters :' containing a table with columns 'Name' and 'Parameter'. The table has one row with the number '1' in the first column and empty cells for 'Name' and 'Parameter'. At the bottom are three buttons: 'OK', 'Get Fields', and 'Cancel'.

	Name	Parameter
1		

## 5.6. Sorted Merge

The sorted merge step merges rows coming from multiple input steps providing these rows are sorted themselves on the given key fields.



The screenshot shows the 'Sorted Merge' dialog box. It has a title bar with a blue background and standard window controls. Inside, there is a 'Step name' text input field with the value 'Sorted Merge'. Below it is a section labeled 'Fields :' containing a table with columns 'Fieldname' and 'Ascending'. The table has one row with the number '1' in the first column and empty cells for 'Fieldname' and 'Ascending'. At the bottom are three buttons: 'OK', 'Get Fields', and 'Cancel'.

	Fieldname	Ascending
1		

## 5.7. Merge join

This step performs a classic merge join between data sets of data coming from 2 input steps.

The join options are: INNER, LEFT OUTER, RIGHT OUTER, FULL OUTER. This step too expects the rows to be sorted on the specified key fields.

**Merge Join**

Step name: Merge Join

First Step:

Second Step:

Join Type: INNER

Keys for 1st step:

	Key field
1	

Get key fields

Keys for 2nd step:

	Key field
1	

Get key fields

OK Cancel

## 5.8. Socket reader & writer

Socket readers and writers are used to transfer data from one server to another over TCP/IP. The primary use for these steps is in-line in a clustering environment. If you want to use these yourself, make sure to synchronize the preparation and start cycles of the transformations between the hosts. (like the clustered transformation does)

The screenshot displays two side-by-side dialog boxes. The left dialog, titled 'Socket reader', contains the following fields: 'Step name' with the value 'Socket reader', 'Hostname' (empty), 'Port' (empty), 'Buffer size' with the value '3000', and a 'Compress data' checkbox that is checked. The right dialog, titled 'Socket writer', contains the following fields: 'Step name' with the value 'Socket writer', 'Port' (empty), 'Buffer size' with the value '2000', 'Flush interval' with the value '5000', and a 'Compress data' checkbox that is checked. Both dialogs feature 'OK' and 'Cancel' buttons at the bottom.



## 6. Source code improvements

### 6.1. A few extra lines of code

Version 2.1.4 contains 160,000 lines of code.

Version 2.2.2 contains 177,450 lines of code, an increase of 17,450 lines.

Version 2.3.0 contains 213,489 lines of code, an increase of 36,039 lines.

Version 2.4.0 contains 256,030 lines of code, an increase of 42,541 lines.

### 6.2. Trackers

In the period between the release of version 2.3.0 on June 27<sup>th</sup> 2006 until January 25<sup>th</sup>, 2007 we fixed 262 bugs and 51 feature requests where implemented.

### 6.3. Committers

ID / Commits	e-Mail	Name	Country	Work
sboden 316	Svenboden (at) hotmail.com	Sven Boden	B	First rate bug fixing, translations, unit tests, implemented many a change request and was a big help out on the forum. Sven was all over the place :-)
berarma 167	Bernardo (at) tsolucio.com	Arlandis Bernardo	ES	He and his co-workers translated the software AND manuals in Spanish, bug reporting & many code changes.
jbleuel 118	Jens.bleuel (at) proratio.de	Jens Bleuel	D	Various bug reports, fixes, German translations and i18n fixes, big help out on the forum.
qinhui99 18	Qinhui99 (at) hotmail.com	Tom Qin	CH	Translation into Simplified Chinese Various bug fixes
Biswapesh 20	Biswapesh (at) gmail.com	Biswapesh Chattopadhyay	UK	Profiling, sort improvements, stream lookup stress testing, memory consumption analysis, merge join step & many suggestions, bug reports, etc.
Mjansen 3	Mjansen (a) betterbe.com	Michel Jansen	NL	Wrote & maintains paper: "Building data warehouses using open source technologies"
Jdixon 27	Jdixon (a) pentaho.org	James 'Chief Geek' Dixon	US	Added support for JNDI database connections, added checks for nulls in quoting algorithms.
Sven.thiergen 12	s.thiergen (a) itcampus.de	Sven Thiergen	D	Various bug fixes, 2 new sample transformations, German i18n fixes and Props sorting "fun bucket" item.
Wconroy 6	wconroy (at) gmail.com	Bo Conroy	US	FTP/enterprisedt upgrade, console logging improvements
MattCasters	Mcasters (at) pentaho.org	Matt Casters	B	The rest
<b>TOTAL 4,000</b>	<b>10 developers</b>		<b>7 countries</b>	

## 6.4. Bug reporters

ACE., ackdesha, asioli, audincham, begunrom, berarma, bicycler, bmmadsen, btalbot, ChrisPerrin, clavigne, cpitis, dawe, dhartford, dmoran, eddietam, equake, etdube, fjsfjsfs, -furet-, gbacskai, gbottazzi, giyoram, guido.leenders (at) invantive.com, guzaldon, HawkNewton, hazmatt, hoezx6r, huangfan575, illum\_user, itzik, JasonToyne, jbleuel, jdixon, jgargus, Jgunderson, jjkarppi, joebordes, JonathonC, Jonhoni, jason, JuliaNiuNiu, kandrews, lmeader, marc\_swingler, MattCasters, ndefontenay, ngoodman (at) bayontechnologies.com, nricheton, phancox, plissak, pstnotpd, radek, rchristy, RPBouman, rsheldon, RutgerDOW, saratchev, sayuso, sboden, scesbron, seckendorff, shassan2, stemey, sven.thiergen, tardifma, ThePerchik, Tihben, TimPigden, tom.gleeson (at) ireland.com, vishalsaha, wceuppens, wconroy, zverina and of-course all the others that reported bugs as *anonymous*.

***Thank you all, 262 times!***

## 6.5. Feature requesters

Avitous, btalbot, Bumbo, gbacskai, guido.leenders (at) invantive.com, hazmatt, illum\_user, jbleuel, jhaile, kandrews, lmeader, m\_sharp, marc\_swingler, MattCasters, ngoodman (at) bayontechnologies.com, RoelVerbeeck, RPBouman, saratchev, sboden, shassan2, surenm, wconroy and again *anonymous*.

***Thank you all 51 times for the many good suggestions!***

## 6.6. Other contributors

*Disclaimer: The list below is by no means complete. Many people files bug fixes and feature requests anonymously. Others were probably shamelessly bypassed. Please remember that this is not done on purpose but rather a result of the limited time we can spend on these type of documents. If you were one of the victims of this horrible policy, simply contact the maintainer of this document and you will be added immediately. The same is true for everyone that DOESN'T want to be on this list: just let us know.*

- The entire crew at Pentaho for the many suggestions for improvements, bug reports and continued support of the entire Kettle project. Special thanks goes to Gretchen Moran for her successful Kettle Forum migration & the fantastic new forum software.
- Special thanks to Martin Lange from the company Proconis (<http://www.proconis.de>) for writing and donating the the new Javascript (Mod) step.
- Special thanks to Youssef Mrabet for writing and donating the Streaming XML input step.
- Special thanks to Michel Jansen from the company Better.be (<http://www.betterbe.com>) for writing & maintaining the paper: "Building data warehouses using open source technologies"
- Special thanks goes to the entire team behind Bernardo Arlandis at Tsolucio (<http://www.tsolucio.com>)