



Pentaho Data Integration version 2.5.0

Getting the job done

Key differences with Kettle 2.4.0

List of changes on 6-04-'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. Advanced error handling.....	5
2.2. Apache VFS support.....	7
2.3. Re-design of the left tree.....	9
2.4. Databases.....	10
2.5. Repository improvements.....	11
3. Spoon.....	12
3.1. Extra options.....	12
3.2. Mixing rows : trap detector.....	13
4. Steps.....	14
4.1. New steps.....	14
4.1.1. Abort step.....	14
4.2. Changed steps.....	14
4.2.1. Caching slowly changing dimensions.....	14
4.2.2. Modified Java Script Value.....	15
4.2.3. ???.....	15
5. Job entries.....	16
5.1. New job entries.....	16
5.1.1. Create file.....	16
5.1.2. Delete file.....	16
5.1.3. Wait for file.....	16
5.1.4. Put a file with SFTP.....	17
5.1.5. File compare.....	17
5.1.6. BylkLoad into MySQL.....	18
5.1.7. Display MsgBox Info.....	18
5.1.8. Wait for	19
5.1.9. Zip file.....	19
5.1.10. XSL Transformation	20
5.1.11. Bulk from MySQL to file.....	20
5.1.12. Abort job.....	21
5.1.13. Get mails from POP.....	21
5.1.14. Ping a host.....	22
5.2. Changed job entries.....	22
5.2.1. ???.....	22
6. Source code improvements.....	23
6.1. A few extra lines of code.....	23
6.2. Committers.....	23
6.3. Bug reporters.....	24
6.4. Feature requesters.....	24
6.5. Other contributors.....	25

1. Changes summary

1.1. Preface

It was only in early February that we released the previous version of Pentaho Data Integration, version 2.4.0. Originally our plan was to just release an updated point release (2.4.1). However, we received so many contributions and added really cool features so we had to go for a major release anyway.

Although we did gain a lot of new job entries to help you do a better “job” in the work flow department, very exciting changes were done in the transformations. As you can see below, advanced error handling and Apache Virtual File System support are great additions to our software.

This document was written as a special “thank you” note to all people involved in the community and to keep everyone informed about the incredible progress we are making.

1.2. Overview

These are the most notable changes that have been made:

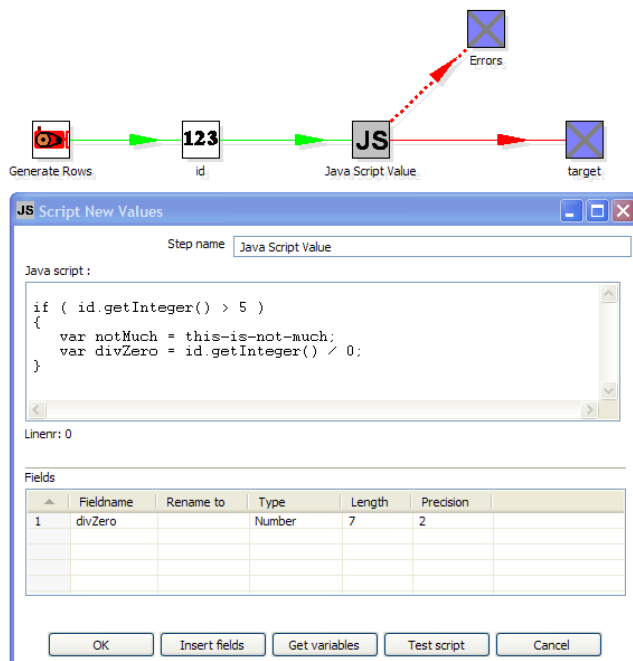
- Advanced error handling added
 - Allowing rows of data that cause an error to be re-routed
 - Allows for cleaner transformations
 - Doesn't slow the transformation down at all
 - Offers interesting new possibilities for quality control and update strategies
- Included Apache VFS support
 - Allows you to read/write files directly from URLs
 - More info is here: <http://jakarta.apache.org/commons/vfs/>
- Re-designed the left trees
 - Turned them into an easier to use toolbar
 - Reduces lookup time for steps and objects
 - Added a new Favorite steps section
- New Steps
 - Abort step : halts a transformation in case one or more records enter the step. This can be used in association with the new error handling capabilities
 - Formula (experimental)
 - Web services lookup (experimental)
- New Job entries
 - Create file
 - Delete file
 - Wait for file
 - Put a file with Secure FTP (STFP)
 - File compare
 - Bulk load into MySQL
 - Display MsgBox info
 - Wait for
 - Zip file
 - BulkLoad from MySQL into file
 - Abort job
 - Get mails from POP
 - Ping a host
- Miscellaneous
 - Hundreds of bugs fixed and dozens of change requests implemented
 - Made the Modified Javascript more compatible with the old Javascript engine.
 - Added caching to the Dimension Lookup/Update step
 - Lots of internationalization efforts took place, for the welcome page, screens and manuals.
 - Easier to create new files
 - Reduced clutter, improved UI usability of Spoon
 - Give various warnings when mixing row layouts at design time
 - >500.000 download attempts of 2.4.0 (including a few DOS attacks)
 - ...

2. General changes

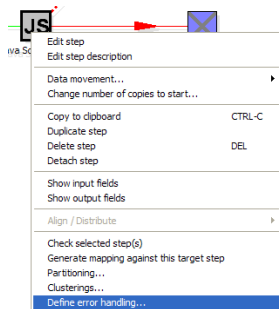
2.1. Advanced error handling

This feature originated as a great idea from Sven Boden, one of the core developers of Pentaho Data Integration. The idea was simple: in stead of halting a transformation when an error occurs in a step, you should be able to pass those rows that cause an error to a different step.

In the example below we artificially generate an error in the Script Values step when an ID is higher than 5.

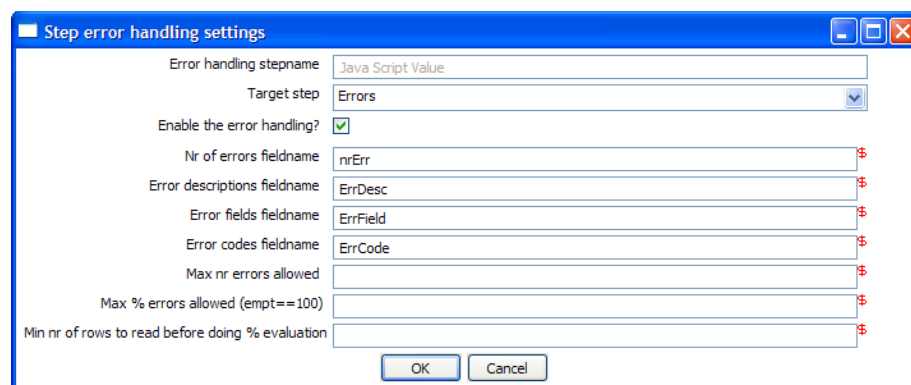


To configure the error handling, you can right click on the step involved and select the “Error handling...” menu item:



NOTE: this menu item only appears when clicking on steps that support the new error handling code.

The error handling dialog looks like this:



Step error handling settings

Error handling stepname: Java Script Value

Target step: Errors

Enable the error handling? ☒

Nr of errors fieldname: nrErr

Error descriptions fieldname: ErrDesc

Error fields fieldname: ErrField

Error codes fieldname: ErrCode

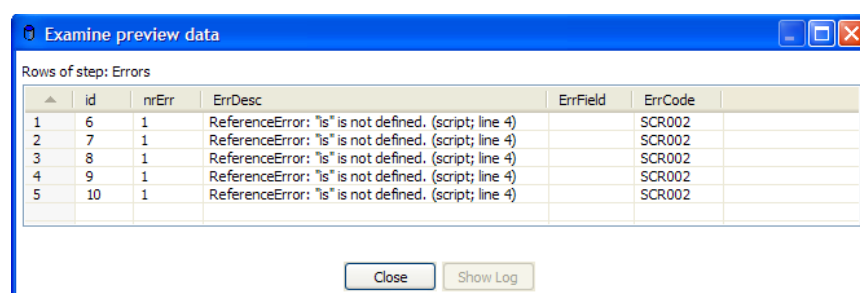
Max nr errors allowed:

Max % errors allowed (empt==100):

Min nr of rows to read before doing % evaluation:

OK Cancel

As you can see, you can add extra fields being to the “error rows”:



Examine preview data

Rows of step: Errors

	id	nrErr	ErrDesc	ErrField	ErrCode
1	6	1	ReferenceError: "is" is not defined. (script; line 4)		SCR002
2	7	1	ReferenceError: "is" is not defined. (script; line 4)		SCR002
3	8	1	ReferenceError: "is" is not defined. (script; line 4)		SCR002
4	9	1	ReferenceError: "is" is not defined. (script; line 4)		SCR002
5	10	1	ReferenceError: "is" is not defined. (script; line 4)		SCR002

Close Show Log

This way, we can easily define new data flows in our transformations. The typical use-case for this is an alternative way of doing an Upsert (Insert/Update):



This transformation performs an insert regardless of the content of the table. If you put a primary key on the ID (in this case the customer ID) the insert into the table cause an error. Because of the error handling we can pass the rows in error to the update step. Preliminary tests have shown this strategy of doing upserts to be 3 times faster in certain situations. (with a low updates to inserts ratio)

2.2. Apache VFS support

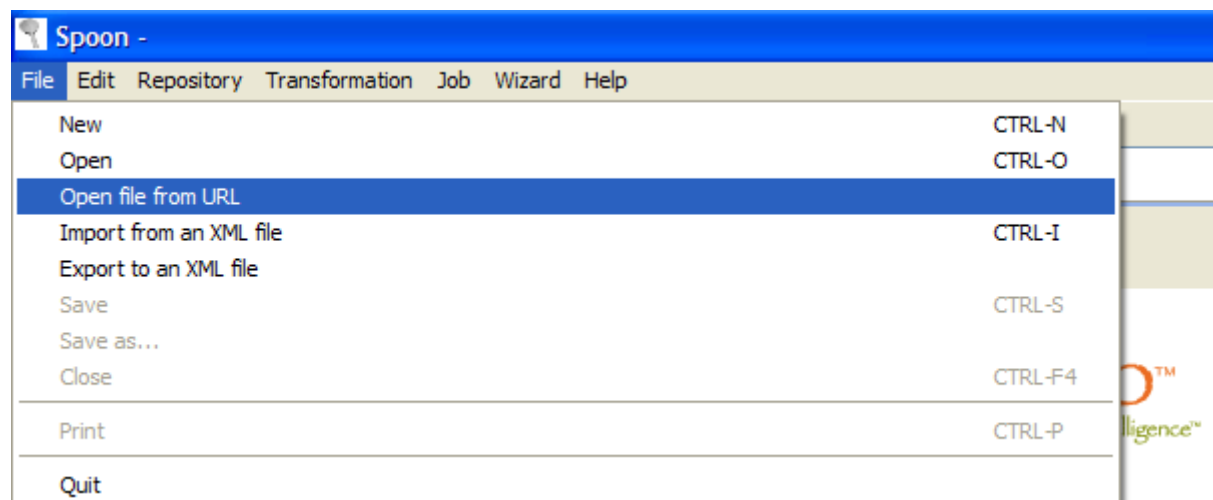
One of the new cool things that we recently implemented is the ability to have reference source files, transformations and jobs from any location you like.

The underlying libraries we use to do that is the [Apache Commons Virtual File System](http://commons.apache.org/vfs/).

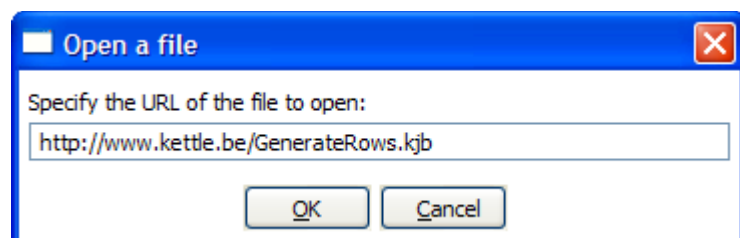
Here is a simple example that you can try with the latest dev version:

```
sh kitchen.sh -file:http://www.kettle.be/GenerateRows.kjb
```

Let's have a look at this job in Spoon. To open it directly from the URL above follow this procedure:



Type in the URL:



Selecting OK will load the job in Spoon:

This job executes a transformation.
The XML for this transformation resides on a webserver in the same directory as this job.
A relative path using internal variables is used to locate the XML file.



The transformation we are about to launch is also located on the webserver. The internal variable for the job name directory is:

Internal.Job.Filename.Directory <http://www.kettle.be/>

This allows us to reference the transformation as follows:

Name of job entry:

Name of transformation: \$

Repository directory: \$

Transformation filename: \$

Please note that if you try this yourself you'll note that you can't save the job back to the webserver. That is not because we don't support that, but because you don't have the permission to so.

Please have a quick look at the almost endless list of possibilities [over here](#). These include direct loading from zip-files, gz-files, jar-files, ram drives, SMB, (s)ftp, (s)http, etc.

We will extend this list even further in the near future with our own drivers for the Pentaho solutions repository and later on for the Kettle repository (something like: psr:// and pdi:// URIs)

As cool examples go, here is one to end with:

File or directory:

Regular Expression:

Selected files:

	File/Directory	Wildcard	Required
1	zip:file:///C:/testfiles/testfiles.zip	.*txt\$	

☐ **Files read**

Files read:

```
zip:file:///C:/testfiles/testfiles.zip!/customer_01_20060801.txt
zip:file:///C:/testfiles/testfiles.zip!/customer_04_20060801.txt
zip:file:///C:/testfiles/testfiles.zip!/customer_02_20060801.txt
zip:file:///C:/testfiles/testfiles.zip!/customer_03_20060801.txt
```

As you can see, you can use a wild-card to directly select files inside of a zip file.

Apache VFS support was implemented in all steps and job entries that are part of the Pentaho Data Integration suite.

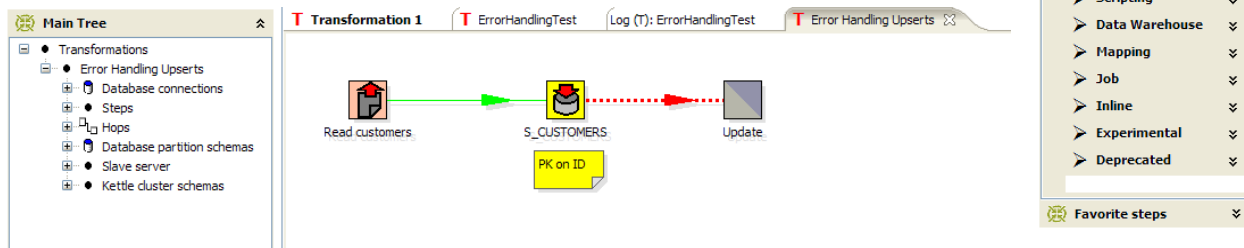
2.3. Re-design of the left tree

The number of job entries and steps keeps growing with every release. Also, more and more people use their own set of plugins and this only increased the number of items in the “Core steps” tree.

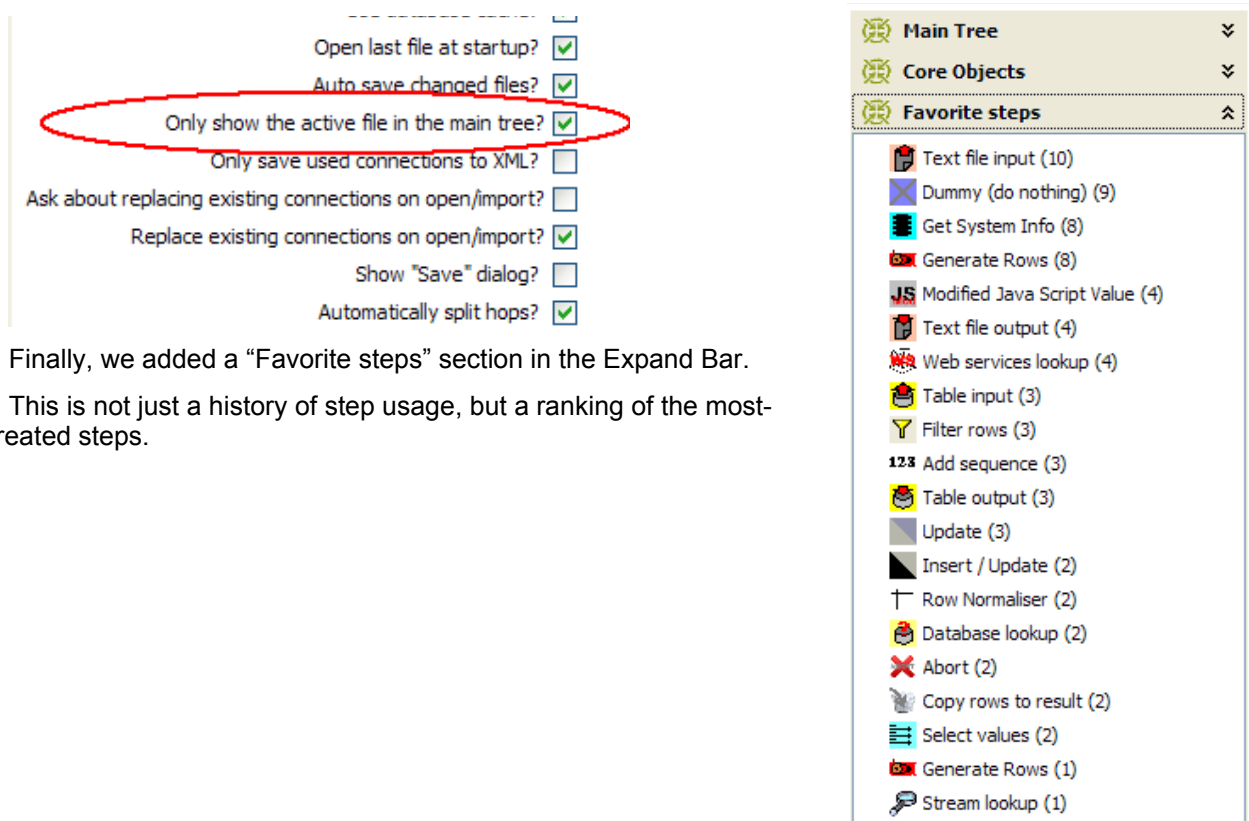
Because of that we looked at an alternative way of displaying these steps. Given the fact that you always only need one step at a time, we went with what is called an “Expand Bar”. You can see an example of this on the right.

The expand bar items only expand one at a time allowing for a less cluttered GUI and easier selection from the range of options.

We also slightly changed the behavior of the “Main tree” on the top left.



The change here is that only the selected transformation or job is displayed in because the tree would otherwise get unwieldy big to the point of being unusable. If for some reason you would still like to see them all, we added an option to influence this behavior:

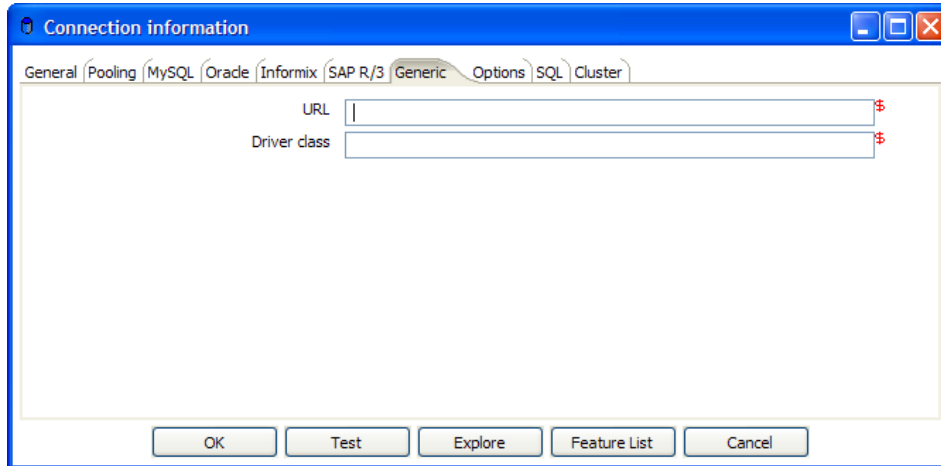


Finally, we added a “Favorite steps” section in the Expand Bar.

This is not just a history of step usage, but a ranking of the most-created steps.

2.4. Databases

- Generic connections can now be defined using variables. That way you can source data from varying database types using the same connection in a single transformation.



- We fixed a nasty bug in the connection pooling authentication mechanism (it actually works now :-))
- Improved quoting of reserved words: when there is a start or end-quote in the tablename or schema, quoting is not done. This allows you to specify the quoting mechanism yourself. Plenty of issues were fixed with regards to schema/table SQL generation and we think that we came a lot closer to an optimal solution now.

2.5. Repository improvements

The repository got slightly changed to allow the definition of short and long descriptions of transformations and jobs.

The image shows the 'Transformation properties' dialog box. It has tabs for 'Transformation', 'Logging', 'Dates', 'Dependencies', 'Miscellaneous', and 'Partitioning'. The 'Transformation' tab is active. It contains the following fields:

- Transformation name: Error Handling Upserts
- Description: (empty text box)
- Extended description: (empty text area)
- Status: Production (dropdown menu)
- Version: (empty text box)
- Directory: / (text box with browse button)
- Created by: - (text box)
- Created at: 2007/04/06 15:34:12.234 (text box)
- Last modified by: - (text box)
- Last modified at: 2007/04/06 15:34:12.234 (text box)

At the bottom are buttons for 'OK', 'SQL', and 'Cancel'.

As you can see, we added a number of fields to the transformation settings:

- Description
- Extended description
- Status (Development / Production)
- Version
- Created by / at

In the repository explorer, this is reflected by the addition of the description field. You can sort on this field as well:

The image shows the 'Repository explorer on [MySQL Localhost]' window. It displays a table of transformations. The 'Description' column is highlighted with a red box.

Name	Type	User	Changed date	Description
Products demo				
PRORATIO				
test				
test copy wizard				
URL-Demo				
Warehouse				
XMLInput				
Zone30				
"a" problem	Transformation	admin	2007/04/06 16:50:59	Test case for the "a" problem in Text File Input
aaaaaaaaaaaa	Transformation	admin	2007/01/18 13:48:30	
Access test	Transformation	admin	2006/10/23 17:10:04	
Alex file 10	Transformation	-	2006/05/15 12:28:04	
Alex file 6	Transformation	-	2006/05/15 12:28:04	
Alex file 9	Transformation	-	2006/05/15 12:28:04	

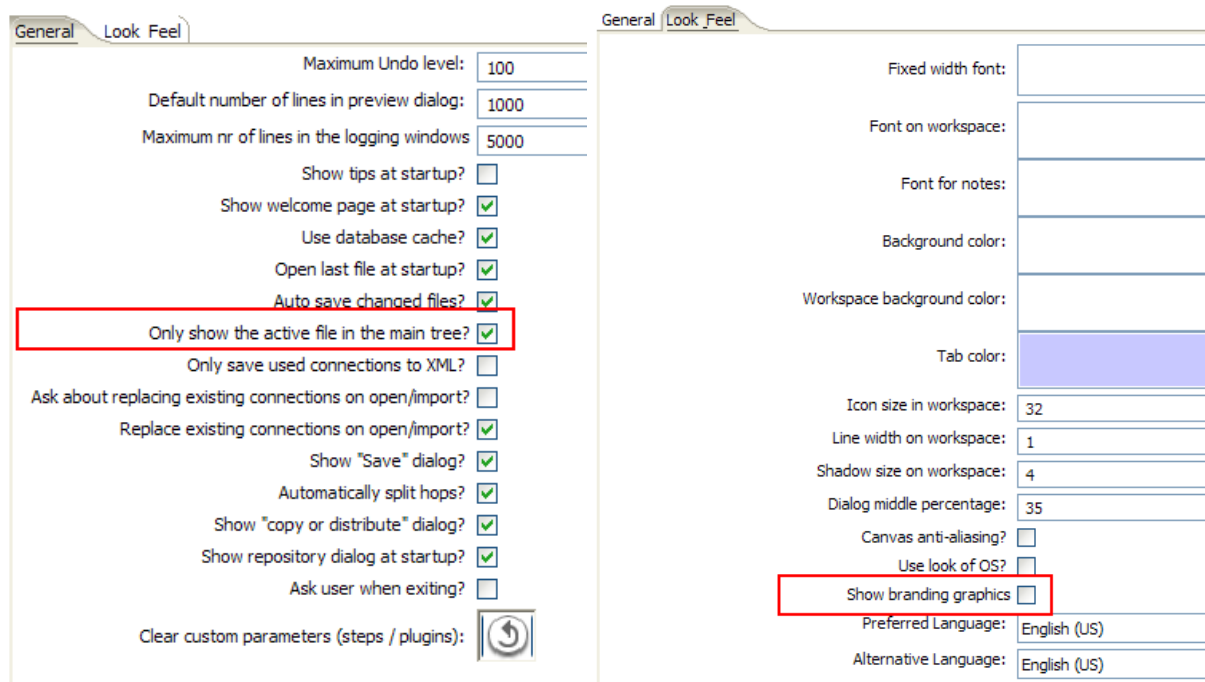
At the bottom are buttons for 'Commit changes' and 'Rollback changes'.

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:



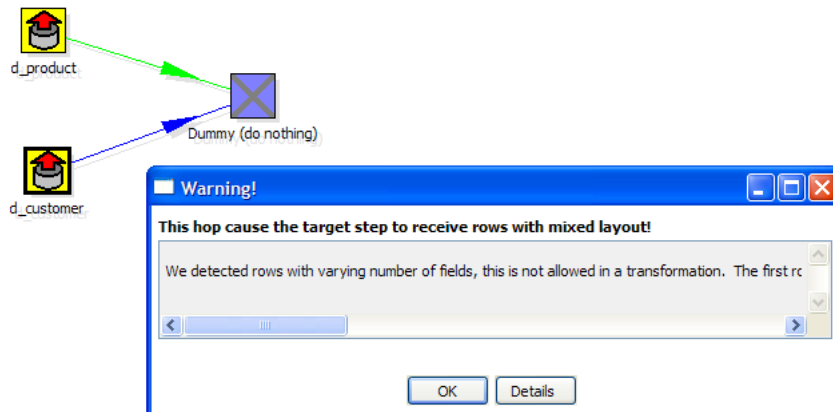
The option on the left is explained above, the "branding graphics" option is displaying some Pentaho Data Integration graphics in Spoon. (for the fans :-))

3.2. Mixing rows : trap detector

Mixing rows with different layout is not allowed in a transformation. However, the developers still receive many bug reports when in fact people are mixing rows of data.

This is causing steps to fail because fields can't be found where expected or the data type changes unexpectedly.

For that reason we added a “trap detector” when you by accident mix rows:



In this case the full error report reads:

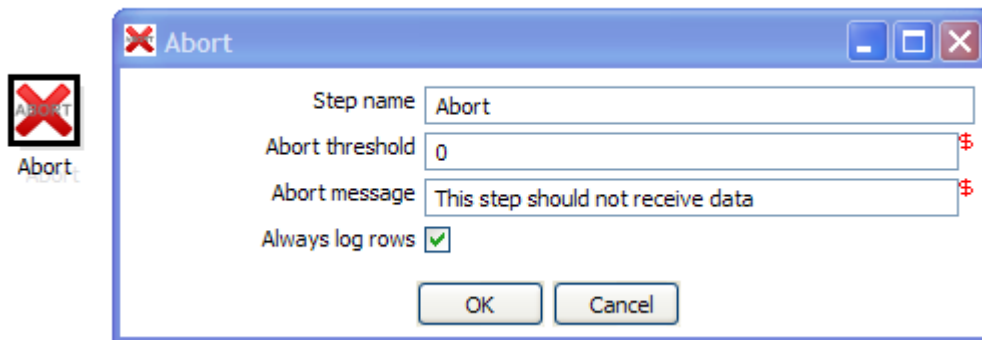
We detected rows with varying number of fields, this is not allowed in a transformation. The first row contained 13 fields, another one contained 16 : [customer_tk=0, version=0, date_from=, date_to=, CUSTOMERNR=0, NAME=, FIRSTNAME=, LANGUAGE=, GENDER=, STREET=, HOUSNR=, BUSNR=, ZIPCODE=, LOCATION=, COUNTRY=, DATE_OF_BIRTH=]

4. Steps

4.1. New steps

4.1.1. Abort step

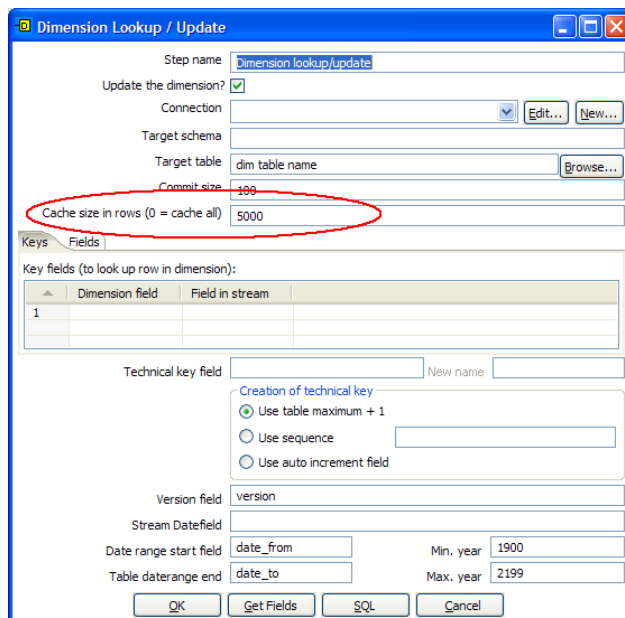
This step can be used in combination with the new error handling functionality. It allows you to abort a transformation when one or more rows are being received.



4.2. Changed steps

4.2.1. Caching slowly changing dimensions

One of the most-requested features was the ability for the “Dimension Lookup/Update” step to cache dimension entries. This was implemented:



The caching works for both the lookup and update mode of the step.

The caching mechanism keeps the highest technical keys in memory for as long as possible because typically those keys have a higher chance of generating a cache hit.

4.2.2. Modified Java Script Value

This step was modified to be as compatible with the older version (Java Script Value) as possible.

4.2.3. ???

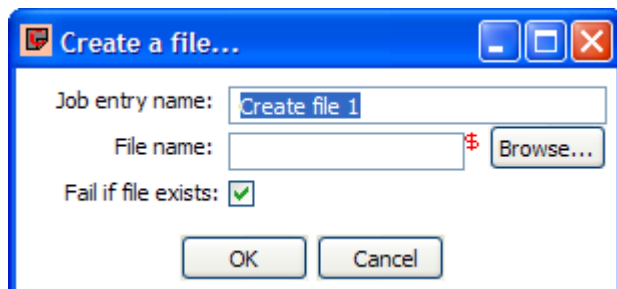
TODO: Check the others..

5. Job entries

5.1. New job entries

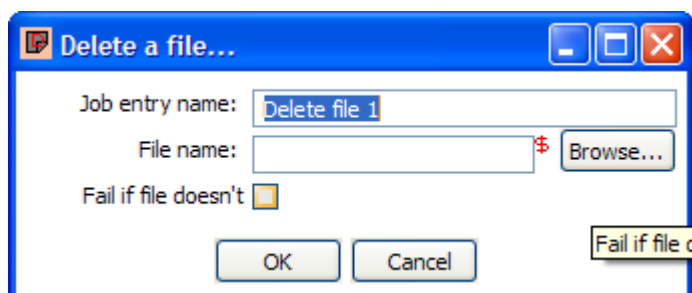
5.1.1. Create file

This is a simple job entry that performs a “touch” (creates an empty file).



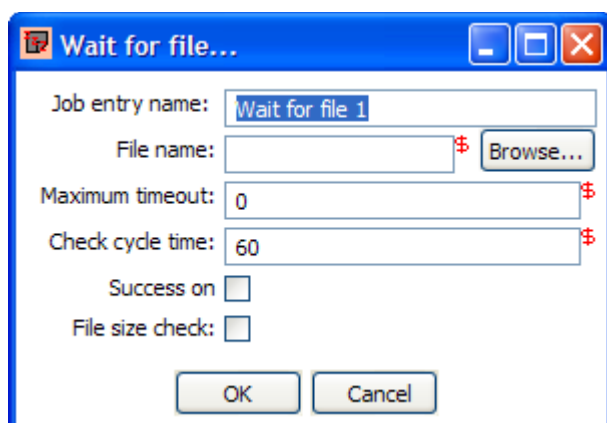
5.1.2. Delete file

This is a simple job entry that deletes an file



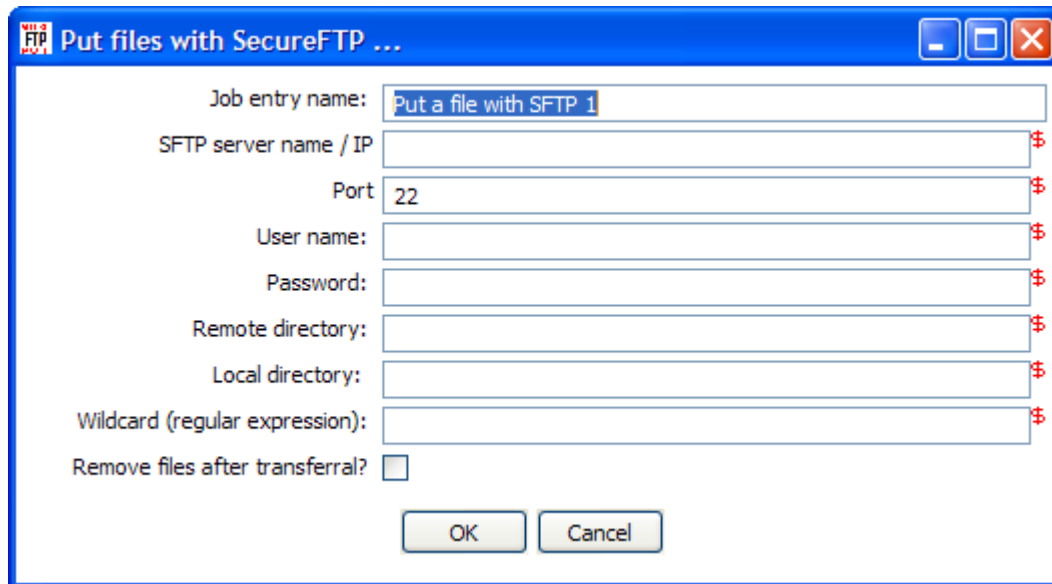
5.1.3. Wait for file

This job entry waits until a file appears.



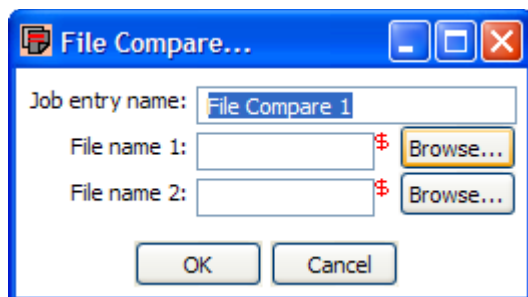
5.1.4. Put a file with SFTP

A job entry to allow you to put files on a secure FTP server.



5.1.5. File compare

If you want to see if the contents of 2 files are identical, this is the job for you!



5.1.6. BylkLoad into MySQL

This job entry uses the MySQL specific “LOAD DATA INTO” SQL command to transfer information from a text to a database table. It has the advantage of being extremely fast.

Mysql Bulk Load ...

Job entry name: BulkLoad into Mysql 1

Database connection: MySQL Local test [New...]

Target schema: [Red \$]

Target table name: [Red \$] [Browse...]

Source File name: [Red \$] [Browse...]

Local: ☒

Priority: NORMAL [v]

Fields terminated by: [Red \$]

Fields enclosed by: [Red \$]

Fields escaped by: [Red \$]

Lines started by: [Red \$]

Lines terminated by: [Red \$]

Fields: [Red \$] [Edit]

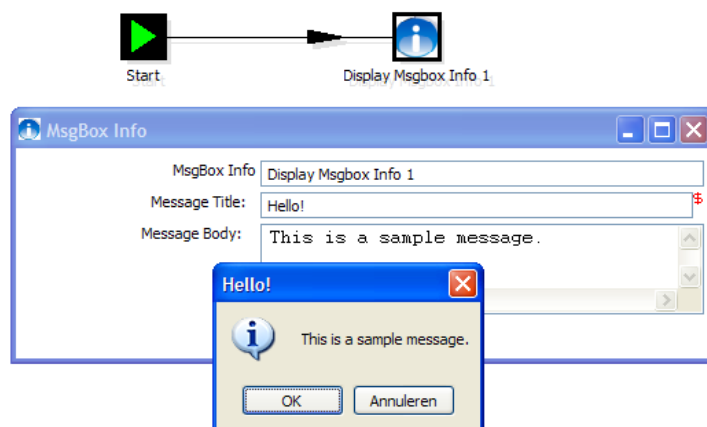
Replace data ☒

Ignore the...first lines: 0 [Red \$]

[OK] [Cancel]

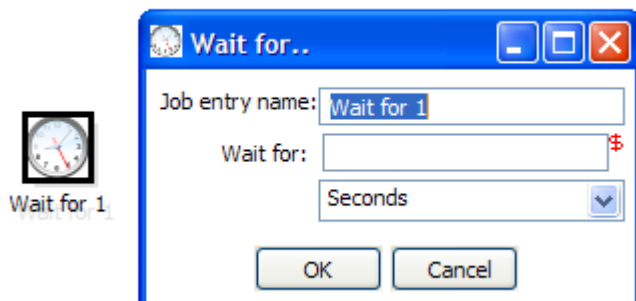
5.1.7. Display MsgBox Info

If you are running in the GUI, you can use this job entry to debug where you are at in the execution of a job. It is **NOT** intended nor designed for batch/runtime use.



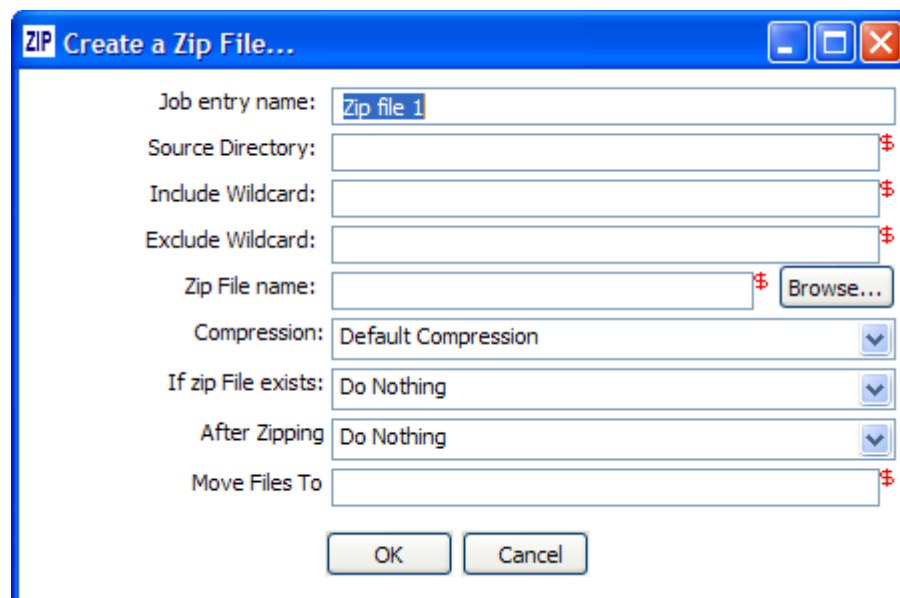
5.1.8. Wait for ...

If you want to put a delay in place before you retry a certain operation, you can use the Wait job entry.



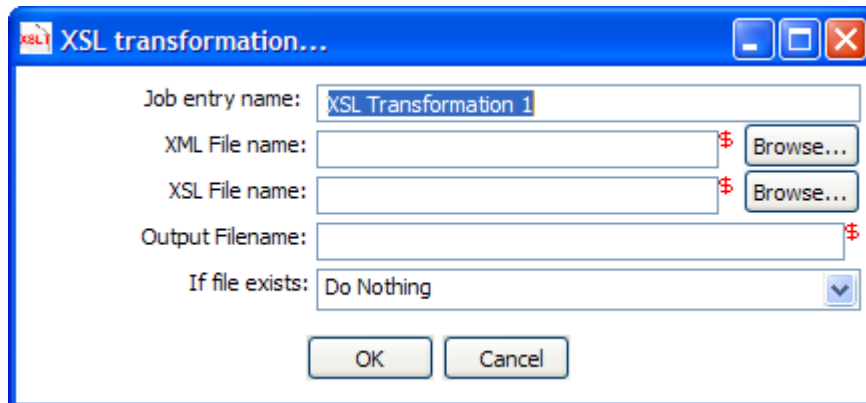
5.1.9. Zip file

If you want to put a number of files into a zip file, you can use this job entry.



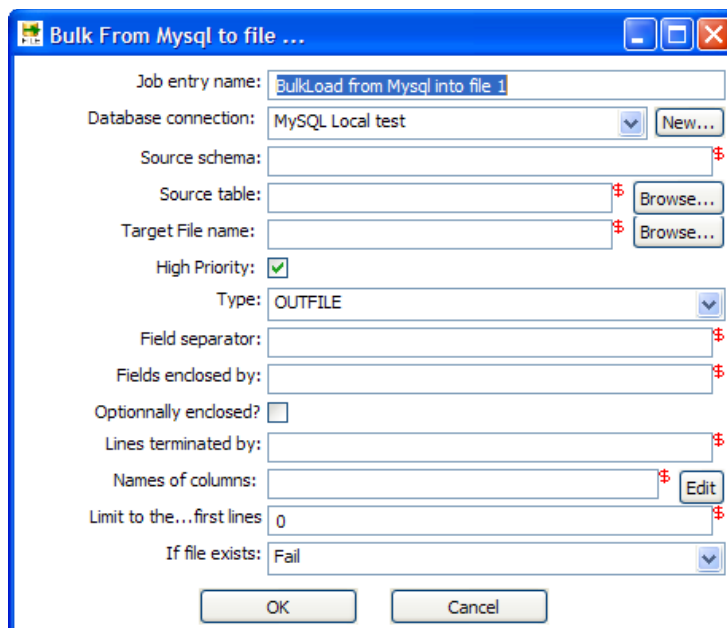
5.1.10. XSL Transformation

If you like to perform XSLT this is the step for you.



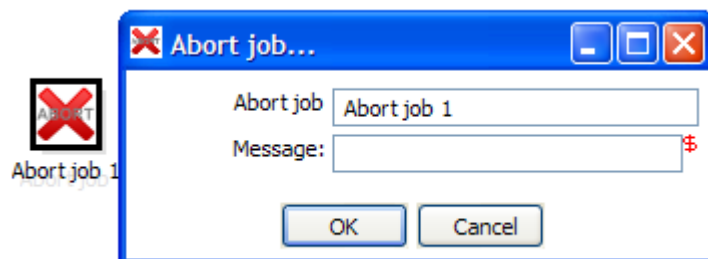
5.1.11. Bulk from MySQL to file

This performs a bulk export from a MySQL table to a flat CSV file.



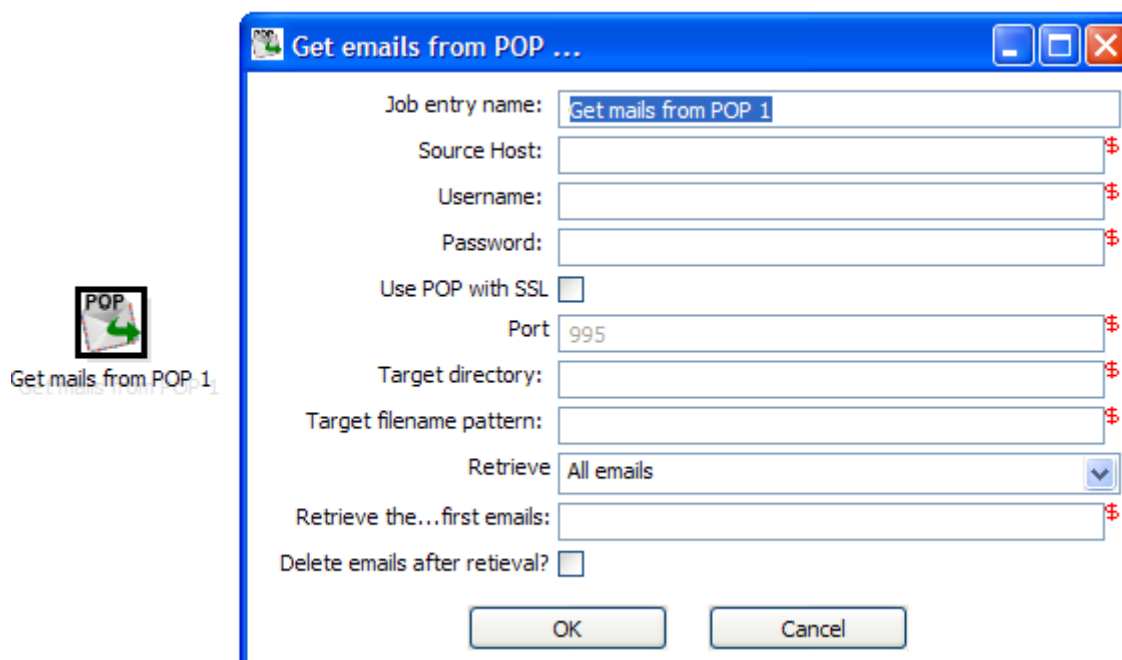
5.1.12. Abort job

If you want to have your job abort execution with an error, you can use this job entry.



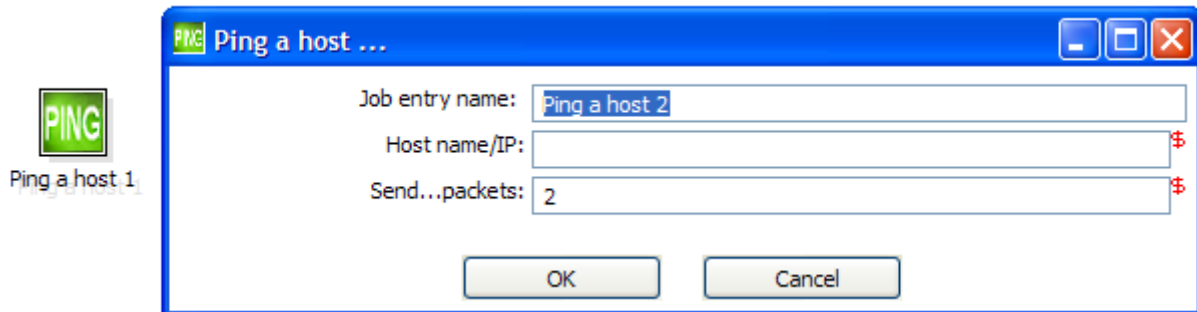
5.1.13. Get mails from POP

This step automatically retrieves e-Mail attachments from a POP server.



5.1.14. Ping a host

With this job entry you can check if a server responds to a ping (ICMP request).



5.2. Changed job entries

5.2.1. ???

TODO: Check each entry for changes.

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.

Version 2.5.0 contains 287,529 lines of code, an increase of 30,499 lines.

6.2. Committers

ID	e-Mail	Name	Country	Work
sboden	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	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	Jens (at) bleuel.com	Jens Bleuel	D	Various bug reports, fixes, German translations and i18n fixes, big help out on the forum.
Sven.thiergen	s.thiergen (a) itcampus.de	Sven Thiergen	D	
MattCasters	Mcasters (at) pentaho.org	Matt Casters	B	A few things here and there :-)

6.3. Bug reporters

Thank you all!

6.4. Feature requesters

Thank you all for the many good suggestions!

6.5. Other contributors

TODO...

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.
- One of the big advances we made in terms of massive parallel processing and partitioning logic couldn't have been made if it wasn't without the help of Google's Biswapesh Chattopadhyay. Biswa had lots of time constraints but he still managed to an awful lot of testing and code writing for which it is difficult to thank him enough.
- 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>)
- Very special thanks go to Samatar Hassan (shassan2) for his enthusiasm, the long list of bugs and change requests filed, continued support and obviously his French i18n efforts.
- Special thanks to the JBoss development team, especially Peter Van Weert (Peter.VanWeert (at) cs.kuleuven.be) and Mark Proctor (mproctor (at) codehaus.org) for donating their implementation of a HashIndex that uses a lot less memory and works faster than the standard Java Hashtable or HashMap classes. It has dramatically reduced memory consumption for the "Stream Lookup" step.
- Herbert Laroca (herbert.laroca (at) gmail.com) for translating into Brazilian Portuguese (pt_BR)
- BreadBoard BI (<http://www.breadboardbi.com>) wrote a nice paper on ETL with Pentaho: http://www.breadboardbi.com/white_papers/pentaho_etl_whitepaper.pdf
- Shibu Mohapatra (shibu_x (at) yahoo.com) & colleagues for the many suggestions.
- Roel VanEck (Roel.VanEck (at) iex.com) for sending patches to improve job entry plugin support. "Group By step" patch and more.
- ... everyone that provided feedback, sent transformations and samples.