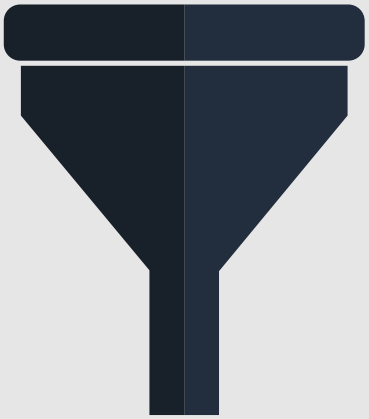


Data and Analytics:
Open Source Data Integration Tool Comparison

excella
CONSULTING

Contact:
Claire Walsh
claire.walsh@excella.com
@datanurturer
703-840-8600

Open Source Data Integration Tool Comparison

Looking for a no-cost data integration (ETL) tool to avoid coding and scripts?

Need more robust data blending capabilities than your BI tool can provide?

This is a scenario we see when we're working at an organization where there is no existing data integration tool and no budget for new tool licenses. Users are getting by using limited data blending functions offered in their BI tool or by taking a crash course in a popular programming language to get the job done (Python anyone?).

There is another option. Community editions of open source, free data integration tools provide organizations with effective, no-cost solutions. In this analysis Excella compares two of the most common tools, Pentaho and Talend, to see how they stack up.

DATA INTEGRATION 101

The dream of having a single source that houses all data for an organization has long passed. Integrating and leveraging data from a number of systems has become a strategic skill for most analysis efforts. Data Integration at Excella is more than simply collecting and merging data together. In many cases, data will be made available at regular intervals and time spent to consolidate and process the data following a standard process can make later blending and analysis easier. There are many aspects considered during the familiar extract, transform and load (ETL) cycle, including:

Data structure design

How does the data need to be stored and presented based upon it's expected use (e.g.; reports, dashboards, extracts, ad hoc queries, etc.)

Data movement

- Designing the optimal path from raw source data to cleansed and enhanced target state
- Leveraging staging data structures where appropriate to enable processing checkpoints and mid-process restartability
- Adhering to tool best practices to create consistent, intuitive solutions that can be transitioned across team members easily

Data quality

- Identifying optimal data quality checkpoints throughout the data processing cycle
- Enabling relevant data quality alerts and reports

Data security & privacy

- Identifying data security requirements to meet federal, state, local and organizational standards
- Leveraging data encryption and obfuscation best practices tailored to meet the need, not one size fits all

WHY USE A DATA INTEGRATION TOOL ANYWAY?

In July 2015, Gartner confirmed the popularity and demand for data integration tools:

…the data integration tool market was worth approximately $2.4 billion in constant currency at the end of 2014, an increase of 6.9% from 2013. The growth rate is above the average for the enterprise software market as a whole…

Here's the 3 key reasons why we advocate using a tool rather than custom code:

1. Solution consistency

In many cases, data integration or ETL scripts are created based upon an individual developer's own coding preferences - their personal best practices and even the choice of programming language. Across multiple team members and over time, this disparity of code can be difficult to maintain causing production support and enhancement work to take longer.

A Data Integration (DI) Tool has a standard interface and while a developer may change the order in which functions are performed, the function screens remain constant across users and solutions. Yes, you need to learn the basic tool functions and once you do, navigation, comprehension and additions to solutions built in the tool are usually much faster and easier than traversing through lines of code.

2. Faster development cycles

Tools provide modules to support common functions. Drag and drop objects to use them, enter the required information, draw a line between objects to create dependencies. In addition, there's usually some immediate quality checks and alerts to let you know when something is missing before you continue building. The experience is visual and intuitive.

By comparison, cutting and pasting code or starting with an existing scripts and modifying them is less instinctive and likely to take longer to update and verify - especially when it's someone else's code.

3. You don't have to be a software developer

Remember when all websites were created by software developers? Nowadays anyone comfortable using a web browser can use a drag and drop interface to create an impressive looking website. Tools exist and persist where there is demand. Drag and drop pre-defined functions allow someone with less experience and/or less technical background to learn faster and become competent sooner. Even if you are an accomplished software developer, the pre-configured functions, automation and real-time alerts a tool provides can save time – and most tools give the option to extend capabilities by executing external scripts or applications when needed.

Business Intelligence vendors appear to be pursuing this vision and are adding data integration functions to create a comprehensive self-service solution for business users – DIY for data. These capabilities are limited so far and data integration is often complex and messy. This is where the pro tools come in.

TOOL COMPARISON MATRIX

Our team of data integration experts evaluated 13 different aspects of Pentaho & Talend to determine which one we think works best.

Here are the links to download Pentaho and Talend.

http://community.pentaho.com/

Scroll down to Downloads section,under Data Integration click the All OS button.

https://www.talend.com/download/talend-open-studio#t4

Click the Download Free Tool button.

pentaho A Hitachi Group Company ⟵ **TOOL COMPARISON MATRIX** ⟶ talend

INSTALLATION PROCESS

Simple and easy to configure environment variables with instructions.

Simple and easy to configure environment variables with instructions.

USER INTERFACE

Spoon

- Jobs and transformations are core building blocks
- Each contain steps
- Steps are linked via jumps indicating dependencies

Easier to navigate the smaller number of options tailored to common integration functions.

TDI

- Projects, jobs and components are core building blocks
- Components bind together via connections
- Jobs are translated into Java or Perl code

Many more details required in user screens and large number of component options can be overwhelming.

TOOL COMPARISON MATRIX (CONT')

USER COMMUNITY

Per Pentaho's website there are 10,000+ production installations and 1500+ commercial customers.

We found answers to our questions quickly online at the following links:

Documentation at Pentaho Community Wiki: http://community.pentaho.com/projects/data-integration/

Ask questions online in the Pentaho Community Forum: http://forums.pentaho.com/forumdisplay.php?135-Pentaho-Data-Integration-Kettle

Per Talend website there are 1300 to 1700 enterprise customers (two numbers published on different parts of the site).

We used the following online links:

User forum https://www.talendforge.org/forum/

LOGGING

- Logs produced at the job and transformation level.
- Able to specific basic to detailed level log options.
- Database logging is limited – look at DBMS log instead.

- Logs produced at the project level.
- Can output logs to database or file.
- Advanced log options to differentiate between statistics logging, metrics logging, process logging.

ERROR HANDLING

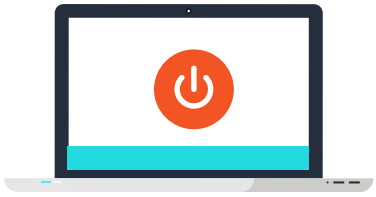
Able to fix issue and continue on from where the job left off.

If job fails have to fix the issue and then restart from beginning.

JOB EXECUTION

To execute a job the options are:

1. Execute directly from Spoon
2. Execute at command level via Pan (for transformations) or Kitchen (for jobs)
3. Carte tool is web server that can execute jobs and transformations
4. Via Pentaho Scheduler or other external schedulers (e.g. cron)



To execute a job the options are:

1. Via tool interface (slow)
2. Export jobs and execute at command line using the .bat or .sh file generated in the export
3. Via Talend Scheduler

TOOL COMPARISON MATRIX (CONT')

PERFORMANCE

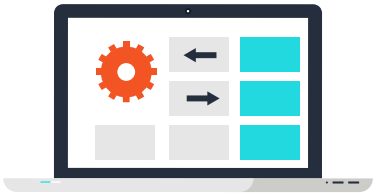
<Note: We worked with small data volumes for this test and are planning a larger scale test in the future>

Slightly faster performance consistently when working in the Spoon interface.

Slightly slower performance when working in TDI.

CUSTOMIZING JOBS

- Limited ability to modify.
- JavaScript code cannot be reused across objects.



- Very flexible.
- All job code generated is visible and modifiable by the user.

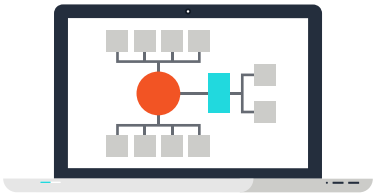
WORK REPOSITORY

- Projects stored as xml files or in a database (for team based sharing).

- Projects stored at the file level.
- Object dependencies are stored in a table and changes made here automatically update all impacted projects.

ARCHITECTURE

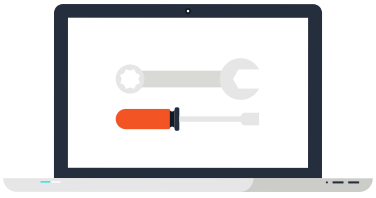
- Multi-threaded Java-based tool, jobs stored as xml files.
- Runs on Windows, Linux and Unix.



- Single threaded code generator (Java or Perl).
- Runs on Windows, Linux and Unix.

TOOL UPDATES

Based upon online feedback Pentaho appears to take longer to add requested changes into new releases.



Talend makes frequent updates and makes new releases available more frequently.

TOOL COMPARISON MATRIX (CONT')

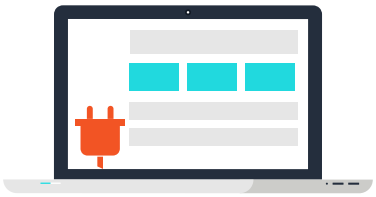
PRODUCT MARKETPLACE

Pentaho Marketplace allows users to explore and test the plugins developed by the Pentaho Community, extending the capabilities of the Pentaho platform.

Talend Exchange provides a market place for contributors and partners to publish Talend add-ons that extend software capabilities.

ADDITIONAL PRODUCT OFFERING

- Pentaho Analytics Platform
- Pentaho Report Designer



Talend Open Studio products for:

1. Big Data
2. MDM
3. Data Quality
4. ESB
5. Talend Data Preparation (self service for text and Excel files)

SIMPLE INTERFACE OR LOTS OF OPTIONS?

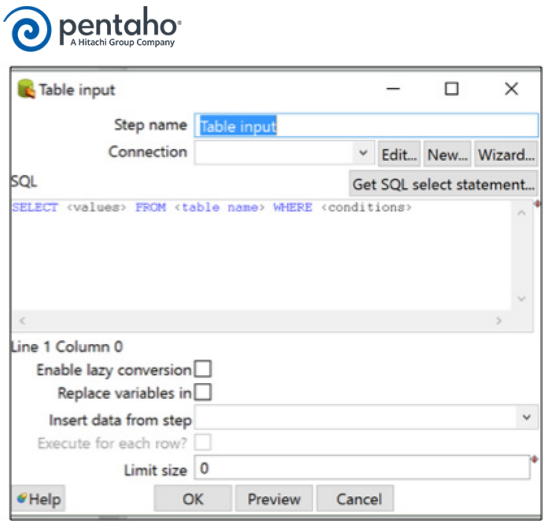
In each tool we tested the following 8 functions:

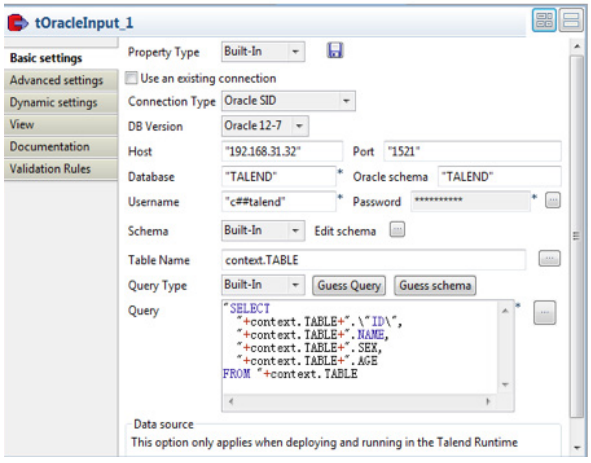
1. Source mapping (CSV, Excel and database table)
2. Sort
3. Group by
4. Filter
5. Table Look-up
6. Dimension Table Update
7. Merge
8. Output mapping (text file and database table)

To illustrate the difference in the tools, below we show screen shots of the same function (using a database table as a data source) in both tools.

The Pentaho screen is more streamlined and takes the wizard-through approach.

Talend asks for a lot more detail and gives many more options (see Talend's array of tabs beyond 'Basic settings').





OUR CONCLUSION

While both tools offer robust functionality for the free, we're leaning towards Pentaho Kettle as our tool of choice. We have junior, mid and senior level team members and we often work with clients who are starting down the path of self-service data integration. Overall, the Pentaho interface (Spoon) is more intuitive and easier to learn, the job performance is slightly faster and we found the user community and online help available to be more helpful than the Talend Community (which is paramount when using a free tool).

That said, Talend offers a far greater ability to customize than Pentaho and if you're comfortable with writing code and have more advanced requirements this may be a better option for you.

We'd love to hear about your experiences with these tools and maybe there's others out there you think we should evaluate. Contact us and tell us what you think.

- **Claire Walsh** - Data & Analytics Lead
- **Brian Rodrigue** - Data Integration Specialty Lead
- **Yadu Mummadi** - Data Integration Developer