ETL Tools Comparison

Written by: Jonathan Levin
March 20, 2008
http://mysqlbarbeque.blogspot.com

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
- Conclusion

Table of Contents

- Introduction
  - What do ETL tools do?
  - Why use an ETL tool?
- ETL tools
- Comparison
- Use Cases
- Conclusion

What do ETLs tool do?

An ETL tool is a tool that:

- Extracts data from various data sources (usually legacy)
- Transforms data
  - from -> being optimized for transaction
    to -> being optimized for reporting and analysis
  - synchronizes the data coming from different databases
  - data cleanses to remove errors
- Loads data into a data warehouse

Table of Contents

- Introduction
  - What do ETL tools do?
  - Why use an ETL tool?
- ETL tools
- Comparison
- Use Cases
- Conclusion

Why use an ETL tool?

ETL tools save time and money when developing a data warehouse by removing the need for “hand-coding”.

“Hand Coding” is still the most common way of integrating data today. It requires hours and hours of development and expertise to create a Business-Intelligence-System.

It is very difficult for data base administrators to connect between different brands of databases without using an external tool.

In the event that databases are altered or new databases need to be integrated, a lot of “hand-coded” work needs to be completely redone.

Table of Contents

- Introduction
- ETL tools
  - Pentaho Kettle
  - Talend
  - Informatica PowerCenter
  - Inaplex Inaport
- Comparison
- Use Cases
- Conclusion

Table of Contents

- Introduction
- ETL tools
  - Pentaho Kettle
  - Talend
  - Informatica PowerCenter
  - Inaplex Inaport
- Comparison
- Use Cases
- Conclusion

ETL Tools
Pentaho Kettle

- Pentaho is a commercial open-source BI suite that has a product called Kettle for data integration.
- It uses an innovative meta-driven approach and has a strong and very easy-to-use GUI
- The company started around 2001
- It has a strong community of 13,500 registered users
- It uses a stand-alone java engine that process the tasks for moving data between many different databases and files

Table of Contents

- Introduction
- ETL tools
  - Pentaho Kettle
  - Talend
  - Informatica PowerCenter
  - Inaplex Inaport
- Comparison
- Use Cases
- Conclusion

ETL Tools
Talend

- Talend is an open-source data integration tool
- It uses a code-generating approach and uses a GUI (implemented in Eclipse RC)
- It started around October 2006
- It has a much smaller community then Pentaho, but is supported by 2 finance companies
- It generates Java code or Perl code which can later be run on a server

Table of Contents

- Introduction
- ETL tools
  - Pentaho Kettle
  - Talend
  - Informatica PowerCenter
  - Inaplex Inaport
- Comparison
- Use Cases
- Conclusion

ETL Tools
Informatica PowerCenter

- Informatica has a very good commercial data integration suite
- It was founded in 1993
- It is the market share leader in data integration (Gartner Dataquest)
- It has 2600 customers. Of those, there are fortune 100 companies, companies listed on the Dow Jones and government organization
- The company's sole focus is data integration
- It has quite a big package for enterprises to integrate their systems, cleanse their data and can connect to a vast number of current and legacy systems

Table of Contents

- Introduction
- ETL tools
  - Pentaho Kettle
  - Talend
  - Informatica PowerCenter
  - Inaplex Inaport
- Comparison
- Use Cases
- Conclusion

ETL Tools
Inaplex Inaport

- Inaplex is a small UK company
- InaPlex is a producer of Customer Data Integration products for mid-market CRM solutions
- Inaplex mainly focuses on providing simple solutions for it's customers to integrate their data into CRM and accounting software like Sage and Goldmine

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
ETL Tool Comparison Chart

	Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
Cost	🟢	🟢	🔴	🟢
Risk	🟡	🟡	🔴	🟢
Ease of Use	🟡	🟢	🟢	🟡
Support	🟡	🟢	🟢	🔴
Deployment	🟢	🟢	🟡	🔴
Speed	🟢	🟢	🟢	🟡
Data Quality	🟡	🟡	🟢	🟢
Monitoring	🟡	🟡	🟢	🔴
Connectivity	🟡	🟢	🟢	🔴

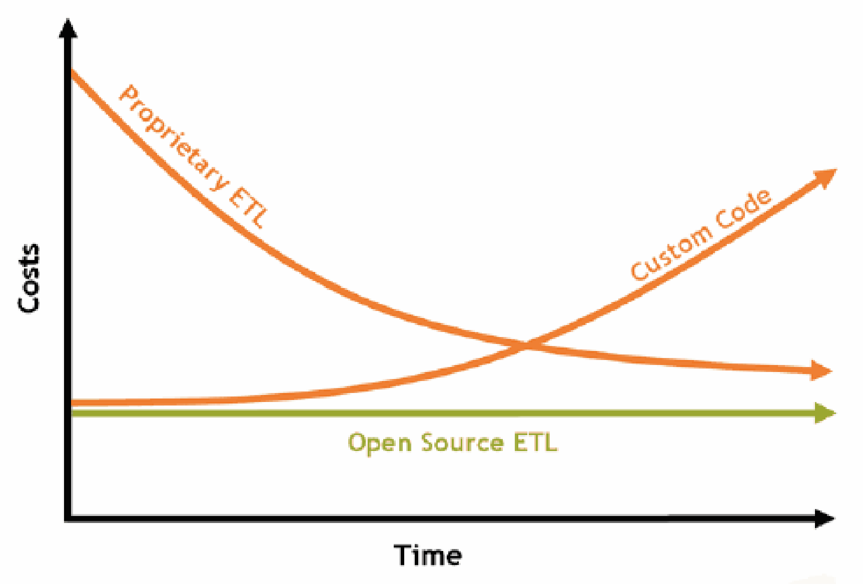
Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Total Cost of Ownership

Total Cost of Ownership means the over all cost for a certain product.

This can mean initial ordering, licensing servicing, support, training, consulting, and any other additional payments that need to be made before the product is in full use.



Commercial Open Source products are typically free to use, but the support, training and consulting are what companies need to pay for.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟢	🟢	🔴	🟢

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Risk

There are always risks with projects, especially big projects.

The risks for projects failing are:

- Going over budget
- Going over schedule
- Not completing the requirements or expectations of the customers

Open Source products have much lower risk then Commercial ones since they do not restrict the use of their products by pricey licenses.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟡	🔴	🟢

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Ease of Use

All of the ETL tools, apart from Inaport, have GUI to simplify the development process. Having a good GUI also reduces the time to train and use the tools.

Talend – Does have a GUI but is an add-on inside Eclipse RC.

Pentaho Kettle – Has the most easy to use GUI out of all the tools. Training can also be found online or within the community.

Informatica PowerCenter – Has an easy to use GUI, but requires some training to make full use of it.

Inaplex Inaport – Does not have a “drag and drop” GUI.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟢	🟢	🟡

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Support

Nowadays, all software products have support and all of the ETL tool providers offer support.

Talend – Offers support, but mainly resides in the US.

Pentaho Kettle – Offers support from US, UK and has a partner consultant in Hong Kong.

Informatica PowerCenter – Offers world-wide support.

Inaplex Inaport – Offers support, but mainly resides in the UK.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟢	🟢	🔴

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Deployment

Talend – Creates a java file or perl file that can be run with an external scheduler on any machine with very little resource.

Recommended one 1Ghz CPU and 512mbs ram

Pentaho Kettle – Is a stand-alone java engine that can run on any machine that can run java. Needs an external scheduler to run automatically.

It can be deployed on many different machines and used as “slave servers” to help with transformation processing.

Recommended one 1Ghz CPU and 512mbs ram

Informatica PowerCenter – Requires a server with platforms: Windows, Solaris, HP-UX, IBM-UX, Redhat, SUSE linux.

Recommended to use two CPUs with 1Gb ram for Standard Edition Server

Inaplex Inaport – Can run on any windows platform that has .NET 2.0 installed

Recommended one CPU with 50mbs ram.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟢	🟢	🟡	🔴

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Speed

The speed of ETL tools depends largely on the data that needs to be transferred over the network and the processing power involved in transforming the data.

Talend – Is slower then Pentaho. It requires manual tweaking and prior knowledge of the specific data source to reduce network traffic and processing.

Pentaho Kettle – Is faster then Talend, but the Java-connector slows it down somewhat. Also requires manual tweaking like Talend. Can be clustered by placed on many machines to reduce network traffic.

Informatica PowerCenter – Is the fastest tool. It has an advanced “PushDown” option that localizes transformation tasks depending on how busy the machine is.

Inaplex Inaport – does not use any special techniques to improve speed.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟢	🟢	🟡

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Data Quality

Data Quality is fast becoming the most important feature in any data integration tool.

Talend – has DQ features in its GUI, allows for customized SQL statements and by using Java.

Pentaho – has DQ features in its GUI, allows for customized SQL statements, by using JavaScript and Regular Expressions. It also has some additional modules after subscribing.

Informatica PowerCenter – does not have that many DQ features, but there is another product called Informatica Data Quality which has many DQ features.

Inaplex Inaport – does have DQ features. Because of the very specific data that Inaport can integrate, it is relatively easy to clean that data.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟡	🟢	🟢

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Monitoring

Monitoring allows to find problems and debug them during and after the development stage.

Talend – has practical monitoring tools and logging.

Pentaho Kettle – has practical monitoring tools and logging.

Informatica PowerCenter – has extensive monitoring tools and logging.

Inaplex Inaport - has practical monitoring tools and logging.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟡	🟢	🔴

Table of Contents

- Introduction
- ETL tools
- Comparison
  - ETL Tools Comparison Chart
  - Total Cost of Ownership
  - Risk
  - Ease of Use
  - Support
  - Deployment
  - Speed
  - Data Quality
  - Monitoring
  - Connectivity
- Use Cases
- Conclusion

Comparison
Connectivity

In most cases, ETL tools transfer data from legacy systems. Their connectivity is very important to the usefulness of the ETL tools.

Talend – Can connect to all the current databases, flat files, xml files, excel files and web services, but is reliant on Java drivers to connect to those data sources.

Pentaho Kettle – Can connect to a very wide variety of databases, flat files, xml files, excel files and web services.

Informatica PowerCenter – Can connect to a huge number of databases, mainframes, flat files, excel files and web services. It can also export as a web service.

Inaplex Inaport – Can connect to any ODBC (windows) connection. It usually gets its data from current databases, outlook, ACT and excel files.

Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
🟡	🟢	🟢	🔴

Comparison
ETL Tool Comparison Chart

	Talend	Pentaho Kettle	Informatica PowerCenter	Inaplex Inaport
Cost	🟢	🟢	🔴	🟢
Risk	🟡	🟡	🔴	🟢
Ease of Use	🟡	🟢	🟢	🟡
Support	🟡	🟢	🟢	🔴
Deployment	🟢	🟢	🟡	🔴
Speed	🟢	🟢	🟢	🟡
Data Quality	🟡	🟡	🟢	🟢
Monitoring	🟡	🟡	🟢	🔴
Connectivity	🟡	🟢	🟢	🔴

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
  - MySQL
  - Loma Linda University Health Care
  - BNSF Logistics
  - U.S. Naval Air Systems Command
- Conclusion

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
  - MySQL
  - Loma Linda University Health Care
  - BNSF Logistics
  - U.S. Naval Air Systems Command
- Conclusion

Use Cases
MySQL

"We selected Pentaho for its ease-of-use. Pentaho addressed many of our requirements -- from reporting and analysis to dashboards, OLAP and ETL, and offered our business users the Excel-based access that they wanted."

Key Challenges

- Reporting and analysis of operational expenses by department and cost center
- Multiple data sources including Microsoft Excel (for cost-center rollups)

Results

- Centralized view of spending by department
- Easy access to information from Excel

Why Pentaho

- Ease of use
- Breadth of solution
- Cost of ownership

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
  - MySQL
  - Loma Linda University Health Care
  - BNSF Logistics
  - U.S. Naval Air Systems Command
- Conclusion

Use Cases
Loma Linda University Health Care

"Pentaho Customer Support has been exceptional. This is a strategic application at LLUHC, and working with Pentaho has accelerated our deployment and improved our overall application delivery."

Key Challenges

- Providing analytics for billing and operations supporting 500,000 patients and 600 doctors

Results

- Comprehensive analysis of time periods, services provided, billing groups, physicians
- Centralized, secured, consistent information delivery (versus prior Excel-based system)
- Ability to drill and analyze down to the individual patient level

Why Pentaho

- Open standards support and ease of integration
- Cost of ownership

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
  - MySQL
  - Loma Linda University Health Care
  - BNSF Logistics
  - U.S. Naval Air Systems Command
- Conclusion

Use Cases
BNSF Logistics

"Using Pentaho for our business intelligence platform, along with the expert support and knowledge provided by OpenBI, BNSF Logistics was able to implement our initial data warehouse with web-based reporting and analytics in just six weeks. Not only did we deliver a powerful business intelligence tool set for our organization in short order, but were able to do so at a fraction of the cost of proprietary alternatives."

Key Challenges

- Cumbersome, manual process for creation and distribution of reports
- Inconsistent data accuracy because of semi-automated preparation processes

Results

- Initial data warehouse with web-based reporting and analytics in 6 weeks
- 75% lower acquisition costs, 50% lower ongoing ownership costs compared to proprietary BI
- Ability to monitor operational business health
- Faster, better decisions in sales processes

Why Pentaho

Open standards support and ease of integration

Cost of ownership

Table of Contents

- Introduction
- ETL tools
- Comparison
- Use Cases
  - MySQL
  - Loma Linda University Health Care
  - BNSF Logistics
  - U.S. Naval Air Systems Command
- Conclusion

Use Cases
U.S. Naval Air Systems Command

"[Open technologies] reduce the cost of software development and they reduce the time in which innovations in software can be incorporated in systems. 'If the project is of a sufficient scale, you cannot get there without an open-source approach,' said Dewey Houck, a senior engineer at Boeing, who spoke at a conference last month about DOD's use of open source." (Government Computer News, Jan. 2008)"

Key Challenges

- Analyzing flight data to reduce operational risk and improve training (human error is a causal factor in 70% of aviation mishaps)

Results

- Ability to leverage recorded electronic sensor data to reduce risk and improve crew performance

Why Pentaho

- Breadth of capabilities
- Proven success and large-scale referenceable deployments
- Successful proof-of-concept
- Dramatically lower costs

Table of Contents

- Introduction
- What do ETL tools do?
- Why use an ETL tools?
- ETL tools
- Comparison
- Use Cases
- Conclusion

Conclusion

- Informatica and Pentaho have very good products.
- Informatica has a far more extensive range of products, but compared to Pentaho is very expensive.
- Pentaho has proved that it can handle small to large scale systems.
- Pentaho is gaining fast momentum with businesses that would not have considered using open source products before.