Microsoft SSIS and Pentaho Kettle:

A Comparative Study for Three-Tier Data Warehouses

Michael L Grecol

Georgia Southern University

Contents

Introduction

Extraction, Transformation, and Loading (ETL) tools integrate heterogeneous schemata, extracting, transforming, cleansing, validating, filtering and loading data from sources into a data warehouse[1]. The ETL process and the associated tools may be used in a greatly varied number of situations where data must be cleaned and moved between data sources. The subject of ETL tools is too broad for a whitepaper; therefore, this whitepaper will only cover the use of two ETL tools (Microsoft SSIS and Pentaho Kettle) within the purview of data warehouse design and implementation.

Microsoft SSIS

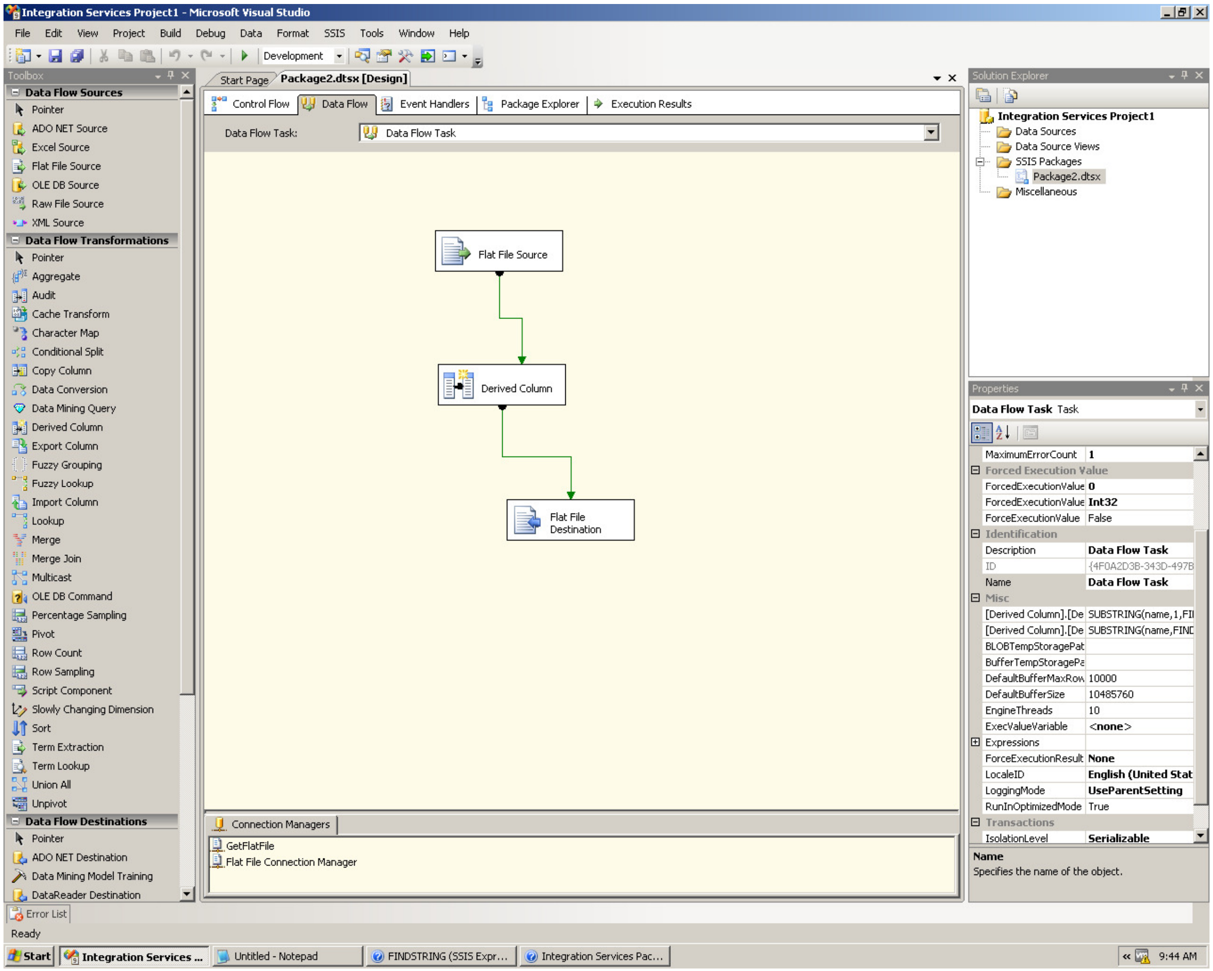
Microsoft provides a proprietary ETL tool named SQL Server Integration Services (SSIS). SSIS is packaged with Microsoft SQL Server and requires a SQL Server License to use it. Microsoft also offers a full business intelligence suite. Additionally, SSIS may be used with a number of database servers through OLE and ADO.NET drivers[2]. Microsoft does not offer the source code as part of the product meaning the developer cannot make modifications to the product to suit the project needs. Also, there is no avenue for a developer to contribute to the future version of the product other than requesting the functionality to Microsoft.

SSIS Development Interface and Capabilities

Microsoft's SSIS system includes three components,

1. Business Intelligence Development Studio (BIDS) seen in Figure 1
2. Data profile Viewer
3. Package Execution Utility

Since most of the development efforts take place in BIDS, the remainder of this paper discusses its capabilities. BIDS is based on Microsoft's Visual Studio IDE. BIDS has a robust drag and drop interfaces and emphasizes configuration over coding. Microsoft documentation claims that you can make robust ETL processes without coding [3]; however, the environment does provide for a script component which allows for scripting in Visual Basic or Visual C#.



MaximumErrorCount	1
Forced Execution Value	
ForcedExecutionValue	0
ForcedExecutionValue	Int32
ForceExecutionValue	False
Identification	
Description	Data Flow Task
ID	{4F0A2D3B-343D-497B
Name	Data Flow Task
Misc	
[Derived Column].[De	SUBSTRING(name,1,FII
[Derived Column].[De	SUBSTRING(name,FIND
BLOBTempStoragePat	
BufferTempStoragePa	
DefaultBufferMaxRow	10000
DefaultBufferSize	10485760
EngineThreads	10
ExecValueVariable	<none>
Expressions	
ForceExecutionResult	None
LocaleID	English (United Stat
LoggingMode	UseParentSetting
RunInOptimizedMode	True
Transactions	
IsolationLevel	Serializable

Name
Specifies the name of the object.

Figure 1: SSIS BIDS Screenshot.

A BIDS project includes data sources, data source views and SSIS packages. The majority of design effort will be spent in the SSIS package area. Each package is comprised of Control Flows, Data Flows and Event Handlers. There are three different types of Control Flows namely:

1. Containers: provide the structure
2. Tasks: provide functionality

3. Precedence constraints: provides connections to and from tasks and containers into an ordered flow.

An example of use of different control flows is in Figure 2. Control Flows allow for execution of various tasks such as SQL, VB, VC# scripting as well as FTP, Send Mail and Web Service Tasks. A total of 28 Control Flow items are available for use. Control Flows may call Data Flows or other Control Flows. The relationship between Data Flows and Control Flows can be seen in Figure 3.

The Data Flows are actual data transformations. Data Flows are divided into Sources, Transformations and Destinations. The Source Data Flows provide components to access a data source using ADO.net, Excel, Flat File, OLE DB, Raw File and XML sources. Data Flow Transformations have 29 transformation tasks to choose from including Fuzzy Lookups, Fuzzy grouping, slowly changing dimensions along with several aggregate functions. SSIS transformation objects are very configurable and favor chaining function calls for string manipulations[3].

SSIS includes robust Event Handlers which can assign any combination of Control Flow tasks for every object in the SSIS project. Twelve events are available for each object such as onPreExecute, onPostExecute and onError [3].

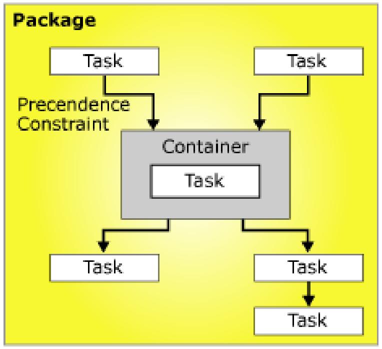


Figure 2: Example Control Flow Tasks.

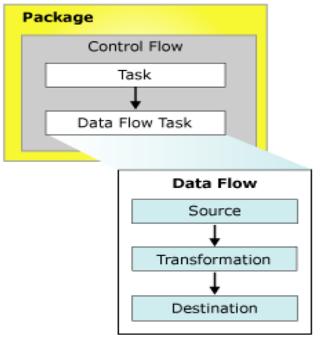


Figure 3:SSIS Control and Data Flow Tasks.

Pentaho Kettle

Pentaho Kettle is an open-source data integration solution. It comes in two versions, The Community Edition, licensed as under the GNU public license and The Enterprise Edition licensed under a commercial license. Pentaho's development community consists of around 8,000 members[4]. Pentaho offers a complete business intelligence suite including data integration, data discovery and exploration and data mining. It connects to any database which can be accessed through a JDBC driver. In addition, Pentaho supports Apache Hadoop and provides an interface into SAP. Since Pentaho source code is available, the program can be modified to meet the needs of the project. Also, the developers can take part in the Pentaho developer community to contribute towards future versions of the product[5].

Kettle Development Interface and Capabilities

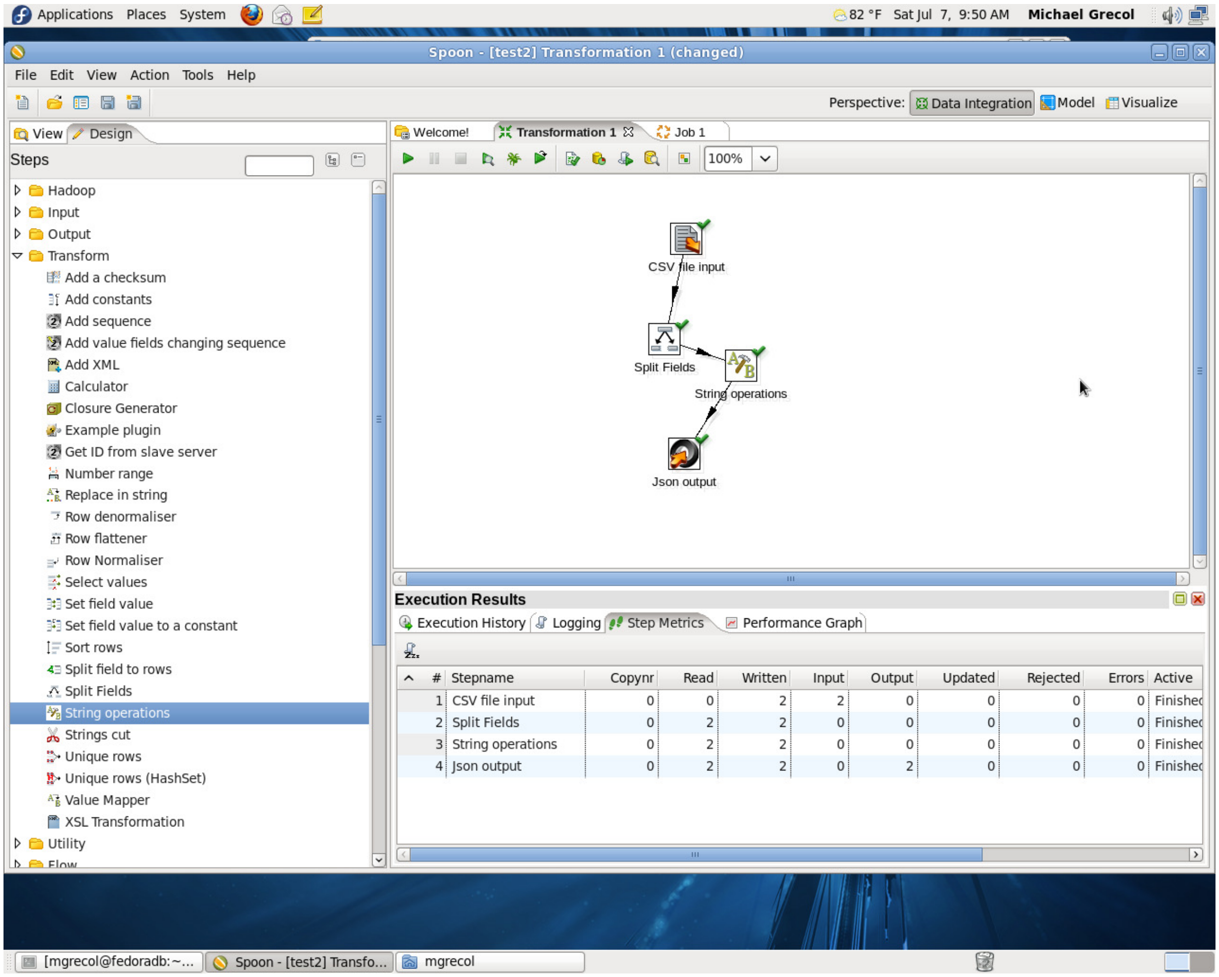
Pentaho Kettle is comprised of four separate programs.

1. Spoon: Pentaho's development environment which is used to design and code transformation jobs.
2. Pan: for running transformation XML files created by Spoon or from a database repository.
3. Kitchen: Kitchen is for running transformation XML files created by Spoon or from a database repository which are scheduled to run in batch mode.
4. Carte: Carte is a web-server which allows execution of transformations remotely.

Because the majority of development effort will take place in the Spoon program, the remainder of this paper will deal with its capabilities. Pentaho's development user interface (Spoon) is based on the Java-based Eclipse IDE as seen in Figure 4. The tool is organized under three viewing perspectives:

- Data integration: Allows for design of ETL processes and jobs through drag and drop, configuration and scripting.
- Model: Allows for designing OLAP metadata models
- Visualize: Allows for testing of OLAP metadata models

Spoon uses a drag and drop interface with a long list of built-in functions. In Spoon a data integration job consists of a Job and Transformation designs. A conceptual model of Spoon's job structure can be seen in Figure 5. Spoon's job scripting capabilities are robust and include three scripting options: JavaScript, Shell and SQL Scripting. Additionally, nine file transfer functions are available including FTP, SFTP, and SSH. Spoon's transformation components are very specific and include: 38 data input source types including common text file formats, office applications, database tables and a direct interface to SAP as well as other applications, 26 individual transformations and 15 Lookup methods including file system operations, database queries and web service calls. String manipulations are simple and to the point favoring regular expressions for validation and transformation. Transformations include seven scripting options such as SQL, JavaScript and Java classes. Spoon also allows the developer to develop or use existing plugins to add functionality[5].



Execution Results

Execution History | Logging | Step Metrics | Performance Graph

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	CSV file input	0	0	2	2	0	0	0	0	Finishec
2	Split Fields	0	2	2	0	0	0	0	0	Finishec
3	String operations	0	2	2	0	0	0	0	0	Finishec
4	Json output	0	2	2	0	2	0	0	0	Finishec

Figure 4: Screenshot of Spoon.

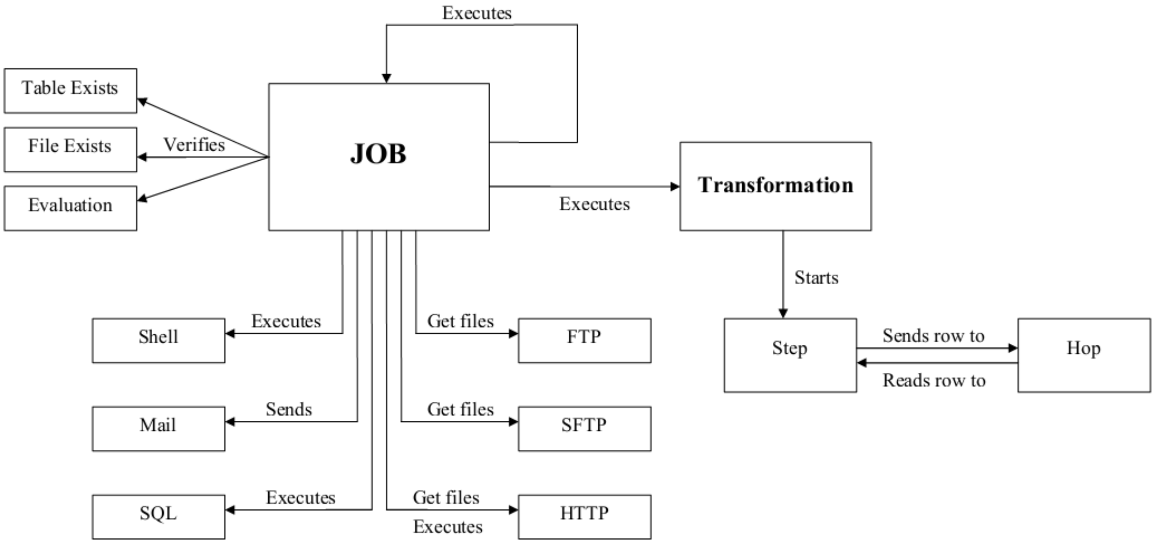


Figure 5: Spoon’s job conceptual model.

A Taxonomy of ETL Activities by Tool

Below is a table showing the built-in transformations provided by both tools. The table concept was provided by [6].

Class	Transformation Category	SQL SSIS	Pentaho Kettle
1:1	Row-Level: Function that can be applied locally to a single row.	- Character Map - Copy Column - Data Conversion - Derived Column - Script Component - OLE DB Command - Cache Transform - Slowly Changing Dimensions - Other Filters (not null, selections, etc)	- Add Checksum - Add Constants - Add Sequence - Add Value fields changing sequence - Add XML - Calculator - Number Range - Replace in String - Select Values - Set field Value to a constant - Split Fields - String Operations - Strings Cut - Value Mapper - If field Value is null - Null if... - ETL Metadata injection - Filter Rows, Last Row, Java Filter, Regex Evaluation - Scripting, Java, JavaScript, SQL.
N:1	Unary Grouper: Transform a set of rows to a single row.	- Aggregate - Pivot	- Row Flattener - Unique Rows - Unique Rows (HashSet) - Analytic Query - Group by - Memory Group by - Univariate Statistics
1:N	Unary Splitter: Split a single row to a set of rows.	-Unpivot -Fuzzy Grouping	-Row Normaliser - Split Fields to Rows - Clone Row
N:M	Unary Holistic: Perform a transformation to the entire dataset(blocking).	- Sort - Percentage Sampling - Row Sampling	- Sort Rows - XSL Transformation - Change file encoding - Sample Rows

	Binary or N-ary: Combine many inputs into one output.	Union Like - Union All - Merge Join-like - Merge Join - Lookup - Import Column - Fuzzy Lookup - Term Extraction - Term Lookup	Join-like - Get ID from Slave Server - Row Denormaliser - Set Field Value - Append streams - Database Join - Database Lookup - HTTP Post, client, REST, Stream, SOAP Lookup. - Sorted Merge - Merge Join - Merge Rows (diff) Union Like - Join Rows
	Routers: Locally decide for each row, which of the many outputs it should be sent to.	- Conditional Split - Multicast	- Process Files - Switch/Case - Dynamic SQL Row - Mapping (input, output, sub transformation)

Table 1: Taxonomy of ETL activities by tool [3, 5, 6].

ETL Processes

The ETL process takes place when the data warehouse is first populated and during every update cycle[1]. Being an important part of a data warehouse project it is typical that implementing the ETL process may be the task consuming the greatest effort[7]. The ETL process is defined as incorporating the following steps [1, 8, 9]:

1. The Extraction Phase: extraction of information.
2. The cleansing Phase: scrubbing of the subject dataset.
3. The Transformation Phase: transforming data to the format of the data warehouse.
4. The Loading Phase: data is loaded into the data warehouse.

There are a large number of ETL tools available which can execute these tasks. There is no set standard regarding how to develop the above steps; therefore, each tool has its own set of methods, languages rules and limitations to achieve the same goal. This complicates the selection of an ETL tool as it is difficult to do apples to apples comparisons. This paper compares two products which are very much are opposites in terms of development philosophy, Pentaho Kettle and Microsoft SSIS and compares their capabilities to perform each phase of the ETL process.

Extraction Phase

The Extraction Phase centers around the tools ability to retrieve relevant data from heterogeneous sources[1]. SSIS offers options to extract raw files, flat files, XML files, Excel as well as ADO.net and OLE DB sources. It's 1:1, Binary and N-ary transformation tools are sufficient to extract the needed data. Each tool offers ample configuration abilities. The event handlers give the developer a strong tool to control the flow of SSIS packages. Pentaho Kettle offers a longer list of available tools. Kettle's tools are designed for more specific purposes; therefore, easier to use. However, Kettle does not offer the robust event model that SSIS does. On the other hand, Kettle offers many more data source type options than SSIS including interfaces to Apache Hadoop, SAP and Google Analytics as well as decryption options. Lastly, Kettle offers the flexibility of Shell scripting and JavaScript to increase its overall capabilities.

Data Cleansing Phase

The Data Cleansing Phase is crucial to the data warehouse system because it improves the data quality[1]. SSIS offers lookup features such as OLE lookup and fuzzy lookup which can help improve data quality. However the pre-made tools fall short for complex data cleansing. As in The Extraction Phase, SSIS's event model offers the developer an added layer of control and execution options. Kettle offers much more cleansing options with the ability to lookup using values using more types of web services and file types. Additionally, Kettle's use of Regex, Custom Java Classes and JavaScript offer the developer simple but powerful tools to validate and cleanse data. Microsoft's claims of code-free ETL design may be overstated; to achieve complex data cleansing code will have to be written in C# or if data is appropriately staged in SQL Server, T-SQL.

Transformation Phase

The Transformation Phase converts data from the operational source format to the data warehouse format[1]. Golfarelli and Rizzi define main transformation processes as conversion and normalization along with matching and selection operations[1]. SSIS offers processes to transform data into destination formats using transform and lookup functions. Character and math functions are provided through chaining standard functions. Basic lookups are provided as well as cache transform and deriving columns. Kettle's offers many more transformation and lookup functions than SSIS. Each Kettle function is designed for a specific use case. Kettle offers a large array of row-level functions and unary groupers which are well suited to the transformation phase of the ETL process. Additionally, powerful JavaScript, user-defined Java classes and custom plug-ins increase its flexibility. If the developer needs functionality which doesn't exist in the program the source code is available and can be modified to suit the needs of an enterprise.

Loading Phase

The Loading Phase is the last step in the ETL process and consists of two methods, either refresh or update[1]. SSIS is well suited to load into many types of databases through its ADO.net and OLE DB objects. Additionally, it has a dimension processing object which allow the developer to define the refresh or update methods. Additionally, SSIS offers a partition processing object which works in conjunction with SQL Server's Analysis Services partitions. SSIS offers very limited output file types. Kettle allows loading through any database which has a JDBC driver. Additionally, Kettle offers a number of output types to be used in The Loading Phase including XML, SQL file output, Access and Excel. In addition, Kettle offers bulk loading for many databases such as MySQL, MS SQL and Terradata and others. Kettle also has an interface to Apache Hadoop built in which adds distributive processing capabilities[10].

Use in a 3-tier Data Warehouse

The chief difference in a three-tier data warehouse is that reconciled operational data is stored in its own database layer[1]. This condition produces new challenges to data warehouse designers and the ETL tools they use. Namely, the transformation layer outputs directly to the reconciled data layer[1]. Additionally, The Loading Phase will be more complex as the reconciled layer and the data marts may be located on different servers. Due to these factors, a robust ETL tool must be used. SSIS is suitable for such use; however a code-free design is unlikely with the complexities that will be encountered. SSIS has a limited number of transformation and loading objects which means a less clean design as developers are forced to get more functionality out of limited tools. SSIS has advanced partitioning features but are limited to SQL Server Analysis Services installations.

Pentaho Kettle is very well suited to the complexities of the three-tier data warehouse. Due to the large number of specialized transformation objects, the resulting design will be more elegant and maintainable. Kettle has bulk load options for several popular databases allowing the enterprise to choose the database that fits the needs of the data warehouse. Additionally, with built-in Hadoop functionality, developers can manage ETL processes across clusters of servers which leads to more options in the overall data warehouse design.

Summary

Both SSIS and Kettle are robust solutions to perform ETL in a three-tier data warehouse. SSIS emphasizes configuration over coding; however, because of limited amount of transformation objects available, coding will be required to process complex data. SSIS's strength comes from its control flow, data flow and event driven architecture. It allows great flexibility to the developer to design the structure and flow the ETL process. Because of limited support for non-Microsoft databases, SSIS is more suitable for the enterprise which predominately uses Microsoft SQL Server and has T-SQL, C# experts on staff.

Pentaho Kettle offers more transformation objects which are more straightforward. It includes many more options to access outside data such as an SAP interface, Google Analytics and several options to access web services. It can be used on either Windows or Linux operating systems. Kettle's strength comes from the ability to use shell scripting, JavaScript, user-defined Java classes, custom programmed plug-ins and the ability to modify source code to meet the needs of the project. Loading is efficient with many bulk loading options for major database servers. Additionally, having Hadoop functionality built in increases the ability of the developer to employ clusters of servers performing parallel processing.

References

[1] M. Golfarelli and S. Rizzi, *Data Warehouse Design, Modern Principles and Methodologies*: McGraw Hill Companies, srl Publishing Group Italia, 2009.

[2] R. Parida and C. Sabotta. (2012, 07/05/12). *SSIS and Data Sources*. Available: http://social.technet.microsoft.com/wiki/contents/articles/1947.ssis-and-data-sources.aspx

[3] (2012, 07/05/12). *SQL Server Integration Services*. Available: http://msdn.microsoft.com/en-us/library/ms141026.aspx

[4] J. Foley. (2008, Startup Of The Week: Pentaho Offers Opens Source BI Alternative. *InformationWeek*. Available: http://www.informationweek.com/news/206904925

[5] (2012, 07/05/2012). *Pentaho, Powerful Analytics Made Easy*. Available: http://www.pentaho.com/

[6] P. Vassiliadis, A. Simitsis, and E. Baikousi, "A taxonomy of ETL activities," presented at the Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP, Hong Kong, China, 2009.

[7] T. A. Majchrzak, T. Jansen, and H. Kuchen, "Efficiency evaluation of open source ETL tools," presented at the Proceedings of the 2011 ACM Symposium on Applied Computing, TaiChung, Taiwan, 2011.

[8] S. Alkis, "Optimizing ETL Processes in Data Warehouses," 2005, pp. 564-575.

[9] (2012, 06/25/12). *ETL Enterprise Data Integration, ETL Tools* [Website]. Available: www.etltools.net

[10] (2012, 07/07/12). *Welcome to Apache™ Hadoop™*. Available: http://hadoop.apache.org/index.pdf