Why All Enterprise Data Integration Products Are Not Equal

William McKnight
McKnight Consulting Group
www.mcknightcg.com
Table of Contents

Data Integration Defined ................................................................. 3
Code Generation vs Black Box .......................................................... 5
Transformations .................................................................................. 5
Open Source ....................................................................................... 6
Big Data Support .................................................................................. 7
Component Library ............................................................................. 7
Agility ................................................................................................. 8
Cost ....................................................................................................... 9
Integration with Data Quality, ESB and Master Data Management .......... 10
Conclusion .......................................................................................... 11
About the Author .................................................................................. 11
Appendix: Modern Data Integration Checklist ..................................... 13
Enterprises far and wide are consuming the significant innovation in the data platform marketplace. We have seen more innovation in data platforms in the past 3 years than in the prior 30 years. The innovation has come from expected (database management systems) and unexpected (enhanced flat-file systems and hash tables in the form of Hadoop and NoSQL systems) areas.

The new requirements enterprises have for their data is unparalleled levels of performance, increased agility and putting in place systems that scale with the workload - the levels of which are more difficult than ever to predict.

Data is seldom confined to one system. Architects may choose to minimize redundancy and utilize data virtualization, but undoubtedly many data elements will flow throughout the architecture, sometimes as-is and sometimes transformed for its new purpose. Being skillful in data integration allows an enterprise to take advantage of the data platform innovations by putting workloads in their absolute best platform. Data integration truly holds the modern enterprise together.

The goal of this paper is to highlight data integration challenges, the choices involved in tool selection, and reasons why Talend Enterprise Data Integration should be considered for your next data integration project.

## Data Integration Defined

*Data integration was defined in 2011 by Anthony Giordano as:*

> "A set of procedures, techniques and technologies used to design and build processes that extract, restructure, move and load data in either operational or analytic data stores either in real-time or batch mode."*

Data integration, clearly, involves moving data. It creates necessary data redundancy. The data integration style\(^2\) can be:

1. Extract, Transform and Load (ETL)
2. Extract, Load and Transform (ELT)
3. Extract, Transform, Load and Transform (ETLT)
4. Extract, Transform, Transform and Load (ETTL)

---

\(^1\) Giordano, 2011. Data Integration Blueprint and Modeling: Techniques for a Scalable and Sustainable Architecture. IBM Press, U.S.

\(^2\) Referred to collectively as ETL
Where there are two transform steps, one of them is dedicated to data quality transformations while the other is “forming” transformations to get the data to work in the destination schema (adding row level metadata, time variance, history persistence, splitting data, filtering data, calculations, lookups, summarizations, etc.).

Regardless of the style, functionally and economically speaking, a data integration tool like Talend Enterprise Data Integration provides advantages over custom coding. Few debate this any longer.

Data integration tools provide:

- Consistent approaches to data integration
- Ease of use over custom code
- Handling of metadata, changed data, time variance and other data integration-specific needs
- Improved productivity and collaboration through graphical tools, wizards and repositories
- Prebuilt connections to source and target systems
- Guaranteed insulation from source and target changes
- Job scheduling
- Job migrations

In the process of moving to a data integration tool, some of the benefits of custom coding can be left behind, e.g.:

- Ease of file-based manipulation
- Integration with legacy code
- Programmer experience
- Low Cost

Why not have the best of both worlds – a tool that generates code - and is economical? Talend Enterprise Data Integration offers significant advantages over competitive data integration products.


**Code Generation vs Black Box**

As with a code generation tool, a “black box” engine-based data integration tool will execute what the user interface programs it to execute. It’s not in the demo-level transformations where this distinction makes a difference. It’s in those few complex transformations that take up a majority of the data integration programming time. Actually seeing the code that the tool is generating provides an immense amount of additional information into the process of understanding and optimizing the integration.

Code can be run on a variety of platforms as well.

* Talend, in particular, optimizes their code for the platform the code will run on, so it runs native to that platform, not interpreted, which can impact performance.

Code generators, given the extensiveness of an open programming language as opposed to the capabilities of a vendor engine, provide innumerable additional and more granular data points for the development of complex transformations in data integration. This flexibility provides efficiencies in data integration work that is not evident until the inevitable complex transformation is coded.

**Transformations**

Talend transformations can be executed in memory or in the database engine, which provides a strong ELT option when a powerful database engine is receiving the data. The ELT option is simple to select in Talend as well since there are ETL job components to select for the data integration.

The source can be a relational database, a file, a NoSQL or Hadoop system. It can come from internal, a third-party or the web itself. The load can be of a data warehouse, one of the many flavors of data mart, an operational data store, a staging area, a NoSQL system, a graph database, a cube, Hadoop or a Master Data Management hub. The target schema could be normalized, dimensional (star or snowflake) or one of the many that consciously or otherwise lands in-between. Ideally the ETL is picking up changed-only data, but on occasion all data is sourced.

As the Giordano definition states, the load can be in either real-time or batch mode. The extract can also be either real-time or batch.
The process may add operating statistics in the form of metadata and, in the event of update, may keep the changed data accessible as well.\(^3\)

The process of transformation could add “derived” data – one of many techniques to bring the target schema closer to its actual usage characteristics. Techniques like this reduce the time the user or analyst has to spend to manipulate data and allows them to get more directly to using the data – what all of data architecture is about. It is specifically the transformation logic where you want the extensiveness of code generation over the black box.

Once these capabilities are covered with procedures, techniques and technologies – and that is saying a lot – the goal of enterprise data integration is to be economic.

Open Source

*With an open source heritage, Talend is the truest data integration offering to the spirit of open source, providing much higher levels of functionality in its free, open source version than commercial data integration products with a limited version in open source.*

These limitations include source connectors like SAP and salesforce, technical connectors like Hadoop and NoSQL systems, row limitations, user limitations and limitations on the profile of company that can access the limited product. True open source allows for experimentation without the risk of capital waste. It also opens up to the developer community for the development of connectors, language support, documentation and translation components.

*Every enterprise should have, or be developing, a strategy to deal with open source software, which is becoming more prevalent and more credible.*

Shutting a shop out of open source limits good options from consideration.

---

\(^3\) In a “slowly changing dimensions”, or other, approach
Big Data Support

While a Hadoop file system is the gathering place for transactional big data for analytics and long-term storage, numerous other forms of NoSQL are available and one or more could be needed for a modern enterprise although for quite different purposes. These NoSQL stores primarily serve an operational purpose – with some exceptions noted. They support the new class of “webscale” applications where usage, and hence data, can expand and contract quickly like an accordion. Indeed, “big data” is much more than Hadoop. It includes NoSQL products, including Graph databases - useful in forming relationship and quickly analyzing connections among nodes.

Talend has connectors for NoSQL\textsuperscript{4} in addition to Hadoop\textsuperscript{5}. Talend and its competition have some way to go in terms of supporting all viable NoSQL data stores and must continue to evolve with the increasing functionality of NoSQL, but the big step of going beyond Hadoop and into operational NoSQL has been taken by Talend.

Component Library

Talend’s Hadoop support is based on the use of HCatalog, the abstraction layer in Apache Hadoop that catalogues the HDFS data. Talend was the first tool integrated with HCatalog and supports Hortonworks, Cloudera, MapR and Greenplum distributions of Hadoop.

To load Hadoop is to select the tHDFSOutput component from the component library, under the Big Data folder and drag it onto the Designer for the destination icon. Settings to enter for this component may include the nominal Hadoop parameters of Hadoop version as well as server name, target folder and target file.

This is the method of operation for using Talend’s large, expanding and searchable component library – select the component, drag it onto the Designer and enter any parameters the component needs.

After the task is completed, Talend shows performance statistics and operational metadata such as rows loaded. Talend also has browse options to read data from Hadoop. The component we use is tHDFSInput. There is also tPigLoad, tPigRow and tPigStoreResult for working with data using Pig. Pig support makes it simple to query

\textsuperscript{4} Cassandra, MongoDB, HBase, Riak, Neo4J, Couchbase, CouchDB

\textsuperscript{5} with Talend Enterprise Big Data
Why All Enterprise Data Integration Products Are Not Equal

the cluster. These are some of the thousands of components available. HBase and Hive components mean Talend Enterprise Data Integration is suitable for any enterprise job and impressive especially for big data.

Agility

Beyond the NoSQL connectors, Talend’s advantage is found in the combination of high productivity and low relative cost.

_We are able to begin integrating data in 10 minutes at a client site with Talend. Its agility starts with getting the product initially downloaded and performing basic functions in the environment._

These functions are then “upsized” to be what is really necessary for data integration in the environment. The essence of agility is the ability to start small, grow incrementally and, over time, meet increasing levels of enterprise need.

The steps are:

1. Download the software and install (unzip, skip license prompt for open source)
2. Name a job
3. Connect to a source by noting its technology from a dropdown of over 400 connector possibilities[^6] and logging in to that source; an icon is created in the Designer Page (Talend’s GUI is an Eclipse add-on)
4. Specifying encoding method, if applicable, and field separator in the source
5. Connecting to the target system
6. Automatically or selectively build the schema in the target if necessary with the Designer Page
7. Connect the source and target with a drag and drop
8. Specify any data transformations required using the GUI[^7]
9. Click the Run tab and run the job; job status can be observed throughout the run

_In a few days, a solid programmer can become proficient in Talend._

[^6]: All connectors are part of the open source version
[^7]: Each transformation type is represented by a representative icon
Of course, as you scale up complexity, you would have more steps. The mapping (step 7) could be more complex by adding data transformations. The job would need to be scheduled, not just run. However, the basics remain as shown.

In the simple example, we were absolved from the need to view, much less manipulate, the Java code that Talend generated for the data integration. Inevitably when discussing Talend with a client, the discussion turns to Java and how much Java skills a shop needs with Talend in order to do the required data integration. It could be zero, but I believe every shop should have available Java skills with Talend, for building complex build routines and components, but the quantity is based on:

- **1. Complexity of Data Integration**
- **2. Performance Demands**

Java resources fall under the category of required resources. Looking at resources as an overall category and having used several tools over the years, for accomplishing a similar sized task, I would rate Talend resource requirements as being among the lowest of all tools. I also believe it is going lower as more functionality keeps being added to the component library.

Tools from larger vendors also have the coding option for scaling beyond what the tool generates natively, and that would be in Java, C or a proprietary transformation language.

Another important aspect of agility is code sharing.

*Code is shared easily throughout the developer community with Talend Forge, where we have found many routines to save time. The community is large and it's likely someone will assist with any problem you encounter.*

**Cost**

I've explained Talend's functional approach to data integration. Talend Enterprise Data Integration can effectively be the data integration tool for all data integration jobs I have come across, including mostly for the Global 2000.

*We have been using Talend as a preferred data integration tool for several years now and find it the most productive data integration tool*

---

8 Talend can also generate code in Perl, MapReduce, Pig Latin and HiveQL
I have used and with an interface preferred by most I have talked to that have used multiple data integration products.

Some other tools can also support all data integration jobs, as long as the connectors are available for the source and target technologies. Though Talend boasts a rich connector set to fixed schemas like SAP and salesforce, even that can be circumvented with lots of detailed work with the other tools. However, leadership dictates that any current comfort and industriousness with other products (or discomfort with open source\(^9\)) not be final deciding factors when it comes to 3x to 5x savings.

For the most basic of configurations, the primary competition to Talend would be minimally $150,000 whereas a paid relationship with Talend would be $30,000 to $50,000. Costs scale from there. Large resource requirements and budgets are not required with Talend.

Integration with Data Quality, ESB and Master Data Management

Data integration can work with a number of related toolsets. A certain level of data quality remediation requires a separate data quality tool. This addresses some of the distinction between “forming transformations” and data quality transformations cited above under DI style.

Data integration tools can handle modest levels of data quality violations, but advanced levels of existing and expected ongoing data quality violations will require a data quality tool.

An enterprise service bus (ESB) is an efficient means of routing data within the environment. Organizations with high levels of data movement and/or real-time needs with that data movement need an ESB.

ESBs facilitate data movement with Master Data Management (MDM) data as well. MDM hubs form, collect and distribute enterprise master data in real-time.

Fortunately, Talend Enterprise Data Integration works within a strong ecosystem of open source Talend data quality, ESB and MDM tools.

\(^9\) Of, if getting the enterprise version of Talend, discomfort that a prominent open source version exists in the vendor’s offering set
In addition to traditional data quality, profiling and matching features, Talend’s big data quality and matching functions can be used in MapReduce projects for computationally intensive tasks. Further productivity benefits are gained since all of these products leverage a common unified platform, i.e. the same graphical tooling, repository, deployment, execution and monitoring environment. Talend is an independent vendor who provides levels of integration at the level of vendors many times over its size.

**Conclusion**

Talend Enterprise Data Integration features enterprise-scale, massively parallel integration for all components of the modern enterprise. It has the strong integration and transformation techniques, deployment options, reusability, user interface (UI), data profiling, and extract and load connectivity that is required.

Talend Enterprise Data Integration is true open source, has the strongest big data support and the agility to be up and running fastest and with extreme performance for initial development.

Featuring integration with Big Data, Data Quality, ESB and Master Data Management, Talend Enterprise Data Integration sits in an ecosystem that meets the need in every organization for a robust, enterprise-ready, fully deployable, and low cost integration tool.

**About the Author**

William McKnight takes corporate information and turns it into a bottom-line producing asset. He’s worked with companies like Dong Energy, France Telecom, Pfizer, Samba Bank, Scotiabank, Teva Pharmaceuticals and Verizon -- 15 of the Global 2000 -- and many others. William is author of "Information Management: Strategies for Gaining a Competitive Advantage with Data."

McKnight Consulting Group focuses on delivering business value and solving business problems utilizing proven, streamlined approaches in information management. His teams have won several best practice competitions for their implementations. He has been helping companies adopt big data solutions. William is a frequent international keynote speaker and trainer. He provides clients with action plans, architectures, strategies, complete programs and vendor-neutral tool selection to manage information.
An Ernst & Young Entrepreneur of the Year Finalist and frequent best practices judge, William is a former Fortune 50 technology executive and database engineer. He can be reached at www.mcknightcg.com or 1-214-514-1444.
Appendix: Modern Data Integration Checklist

Does the product…

ARCHITECTURE

• Fully utilize a MPP environment by using synchronous databases across the nodes?
• Support load balancing across the cluster?
• Have a model that allows developers all over the world to contribute useful code?
• Generate code that will run on different machine specifications?
• Generate code that can be broken down and run on different machines?
• Exchange metadata with business intelligence tools?
• Allow for check-in and check-out of code?

FUNCTIONALITY

• Generate code to conditionally pass the data into multiple tables?
• Utilize memory to store tables for lookups?
• Automatically code for slowly changing dimensions, if desired?
• Have a scheduler that can be coded with workflow?
• Allow for easy understanding of data source and target?
• Impute complex derived data, such as analytics, from the stream?
• Collect changed-only data even when that data is not encoded as such?
• Allow transformations to span records, like avoiding duplicates in a batch?
• Allow for seamless change to ELT processing?

PRODUCTIVITY

• Show you the code that it is executing, for easy debugging?
• Have a modern-styled user interface?
• Catalog code and allow for its unlimited reuse?
• Allow for commenting of the code for advanced documentation?
• Economically use the budget since data integration never sits in isolation of other domains to accomplish a business objective?
CONNECTIVITY

- Allow for native connection to current and future systems of interest to the business such as Hadoop and NoSQL?
- Allow for native connection to current and future systems of interest to the business?
- Exist in a family of related products that can be integrated with, including data quality, ESB and master data management?