Analysis of Data Integration Technologies

Prepared by : Christopher C. Biddle

IdealNet, Inc.

Table of Contents

Introduction

Rapid technology change and innovation is an expected component of modern life. This is self-evident today with the feverish pace of development within such sub markets as smart phones and tablets. It is interesting to note, however, that some technology areas experience a perception within the public mind (and within target business customers specifically) of having achieved a "plateau of maturity" that neither requires nor offers opportunity for continued innovation.

Perhaps surprisingly, one such area is within data integration technologies. It is also impacted by today's plethora of both new and old platforms and systems that consistently increase complexity.

As Gartner stated in their 2010 "Magic Quadrant for Data Integration Tools" report, "Market conditions favor offerings with low costs and rapid time to value, but buyers that see data integration as strategic also seek rich capabilities to fuel their information infrastructure."

Evolution

From the inception of data processing systems through the middle of the 80's decade, integration between systems was always a custom, platform-specific affair. Early Mainframes and Minicomputers relied on proprietary data storage systems via indexed flat files, closed simplistic database structures, and so on.

The situation improved as packaged relational database products were introduced and system/network interoperability improved. APIs were created and exposed for customers so that they could more easily move data into and out of various hardware and software architectures. Many of these offerings were driven initially by large players (both hardware and software) at the time such as IBM®, DEC®, etc. In addition, the rise of pure, enterprise software companies which specialized in high performance databases also contributed to the development of the market at this time. Such companies included Oracle® and Sybase®.

A continual reliance on information systems with key databases combined with an exponential growth of data volume precipitated the development of Extract, Transform, and Load (ETL) software packages. These programs were designed to improve data movement efficiency and speed via proprietary software engines that allowed sophisticated users or technical personnel to define and code data transformation rules. The rules would be executed as the information was moving from one underlying data source to another. With the conceptual rise of centralized "Data Warehouses" among corporate customers, these features took on a new importance and the use of ETL products grew quickly.

It was at this point that the public impression of innovation in this area slowed significantly. Just as individuals and business teams achieved a long-term comfort factor with preferred hardware, operating systems, Enterprise Resource Planning (ERP) systems, or databases, virtually all established customers reached a point of standardization on the use of one, chosen ETL platform.

True innovation in this area, however, has not abated and is the ultimate subject of this study and analysis.

Beyond ETL

On behalf of multiple customers, IdealNet, Inc. has periodically managed a variety of software package selection projects. These, of course, have included data integration technologies. Most recently, we have executed these projects in the area known as Master Data Management (MDM) for customer master data and product master data. To ensure current and comprehensive results for our clients, we expanded the footprint of possible solutions to be examined during these projects and were pleasantly surprised at the emerging variety of new and different approaches that improve on the ubiquitous ETL model.

To formalize this task, we created a comparison template which evaluates product and solution offerings in 59 areas (within 10 specific categories) for comparison. This represents the most common and most difficult segments of the data integration challenges which are faced by companies today. These elements include necessary features and functions and are as follows:

Business Application	Data integration - regular data movement
	Data migration - one time data movement between systems
	Business intelligence
	Data Warehouse
	Data Mart
	Master Data Management
	Virtual Master Data Management
Platform Deployment	On-premise
	Software as a service (SAAS)
	Cloud (Public and/or Private)
Connectivity to Data Sources	Direct database connectivity (insert/update/delete)
	Application program interface connectivity (access only through API)
	Flat file
	Standards
Synchronization	Batch
	Real-time or Near Real-time
	Other
Transformation	String
	Math
	Boolean
	Fuzzy Logic
	Complex, Summarization, Statistical
	Custom transforms
	Other
Data Movement	Bulk data movement
	Data federation for BI deployment or equivalent
	Record level synchronization
Test, Development and Operations Environments	Target user
	Dashboard
	Operations Console
	Workflow
	Security
	Data dictionary - metadata repository
	Shared library of transforms
	Test & development environment
	Production environment
	Reporting
	Graphic
	Disaster recovery

Data Modeling	Connection to data sources
	Discovery of metadata structure in data sources
	Representation of metadata structure in data sources
	Automatic update of metadata in data sources
	Semantic discovery support
	Search of metadata across multiple sources
	Direct access to underlying data - in a useful and navigable format - from metadata
	Ability to model all data sources
	Virtual structures in metadata to facilitate mapping
	Lineage of metadata
	Metadata export
Data Quality and Data Governance	Basic data quality capability
	Data governance by implementation of business rules
Architecture	Standalone
	Networked
Standards	SOA
	JDBC
	ODBC - .NET
	Web Services
	Other

Table 1 – Outline of Comparison Categories and Subcategories

The universe of products and solutions which were examined included the most established ETL vendor, Informatica®, several open source platforms which have gained in popularity over the past several years (including Talend™, Pentaho™, and Apatar™), and the product suite of California-based Queplix®. In certain areas, detailed tests were required in order to properly validate each system, but for the purposes of this white paper all comparative results will be condensed into a meaningful, yet abbreviated narrative. Bear in mind that there are a bewildering number of vendors with offerings which propose to address some or all of the challenges that we have profiled. We specifically chose our subset so that (1) we could focus more deeply on each product, (2) to include one of the leading vendors with great longevity, (3) to include representation from successful vendors in the open source community, and (4) to include at least one vendor who comes at these problems from a significantly different vantage point.

Comparison Categories

Business Application

The areas which were examined for the "Business Application" category represent the primary, known data movement and integration scenarios which exist today across vertical industries for all customers. Some, such as Data Integration and Data Migration have the longest history, which predates all of the products being considered for the study, while others such as Virtual Master Data Management have appeared more recently in part rising out of the baseline capabilities that the software tools themselves deliver. Other than full or partial support of the actual functionality, the other points of comparison in this category (as is the case in certain subsequent categories) cross over into specific methodologies or approaches in combination with the ease of user experience in practice.

As an element to examine, Data Integration can be broken into two distinct items in this category. That is, migration of data between two explicit data sources and also data integration between three or more sources. The primary reason for this distinction is that there are clear differences among the products in the realm of three or more sources, while data migration between two sources is well served by all of them. In fact, moving data between two specific data sources is a lynch pin piece of functionality within the long established ETL tools. Informatica® excels at this since it has been in existence for the longest period and began at a time when data movement was typically always done between only two different data sources. This scenario normally encompasses daily, batch-oriented data movements. It is important to note that we found, in practice, that all of these products can and do support this function. Although, we point out here that Informatica® would be the recommended solution based on its longevity with this feature and the fact that it was originally architected with this in mind.

As we move to three or more sources, however, things get more complicated. While Informatica® could be used for this, since it is designed to be a point-to-point solution, it requires an increased effort to set up and support, whereas the other products evaluated here provide improvements. ETL must be connected from "A to B", then "from B to C" and then "from C to A". The technological solution really does not scale well. Either you manually program all of this, as with a legacy MDM hub powered by ETL, or you put your own data mart in the middle of a bunch of ETL and program it very explicitly. The effort is time consuming, very complex and quite difficult. The open source ETL tools provide some additional functionality in managing the multiple point-to-point connections but only in a very primitive way. In contrast, the Queplix® suite utilizes its metadata layer to fully abstract the underlying data sources which allows for "N" number of simultaneous data migrations. In addition, the Queplix® technology includes one-to-many and many-to-one features which include multiple field updates within one record within which each individual field can be tied to and updated from a variety of different data sources (as opposed to the conventional, full row record to record update). Queplix® architecture is driven by data virtualization which is a very different way of building a data integration engine. Data virtualization products have been around for several years and have been used to build out the data warehouse and support business intelligence source connectivity. Queplix® data virtualization services use a hub and spoke architecture to support a persistent metadata server. This server can uniquely enable automated discovery, visualize the data sources, allow for federation and structure necessary transforms. It is a solid concept in a lightweight and easy-to-use package. Coupled with integrated data quality, it provides differentiation to the Queplix® cloud offering (QueCloud™). It should be noted that Informatica® offers data virtualization in their Data Services product suite but they don't really use the data virtualization technology for their mainstream data integration. It remains very centered on data warehousing and business intelligence activities.

We next focused on the area of "one time" Data Migration. This is required when a legacy or prior version of a system is being replaced with newer software. The historical data from the older system must be migrated or converted into the updated system to maintain continuity for the business which it serves. As with Data Integration between two sources, in this case, the ETL tools are leaders. In fact, all of the ETL tools perform this function well, so the primary consideration between them boils down to price and features (particularly when comparing the universe of open source tools to a product such as Informatica®). Despite all of this, we should point out that the Queplix® suite also support this functionality as well.

Another avenue of data needs within businesses appears within the realm of Business Intelligence (BI) products and platforms. In this case, products and vendors such as Cognos®, SAP® (Business Objects®) and Microstrategy®, provide their own, targeted tools and methods for bringing data into their systems. The overall approach is similar to generic ETL, but is, of course, limited to serve the specific BI software. Outside of this, standard ETL tools may also be used, although this then requires multiple steps to implement correctly including: (1) designing and setting up the ETL scripts to bring the data into a database schema which has been designed to meet the needs of the BI tool, (2) mapping this database schema into the BI tool itself as a source, (3) managing any necessary data transformations within the ETL tool prior to populating the database schema for BI use. The Queplix® suite provides a unique differentiator in this area in that it uses the concept of "Software Blades™" that are pre-built connecting engines for a variety of data sources and targets. Within these connectors, underlying data source security, integrity and rules are automatically encapsulated which prevents business or technical users from violating these conventions.

One backbone of modern IT departments is the use of one or more Data Warehouses as well as Data Marts. These, again, involve the regular movement of large amounts of data into the data structures of these areas from multiple, disparate source systems throughout the enterprise. For volume and performance considerations, these mass updates are typically done during nightly and/or weekend processes. The new data is then available on the following business day for users to deploy in analytical and forecasting capacities. With this process being quite similar to Data Migration, we had similar findings with regards to the solutions. All of the ETL tools handle this well, although great care and expertise must be deployed when using them in high data volume situations as very specific performance tuning and considerations come into play. Each of the ETL products differs in various layers of their technology and therefore one may experience a performance issue in one segment of the process while the others do not and vice-versa. Through the use of a metadata layer, the Queplix® suite of products addresses the issue of performance bottlenecks by minimizing the actual movement of underlying large data sets until it is required. This Queplix® feature merits consideration for prospective customers who are evaluating tools for the movement of data into their Data Warehouses and Marts.

Over the past several years, Master Data Management (MDM) has become a "hot topic" for businesses as they grapple with tasks such as improving data reliability, centralizing key data for use enterprise-wide, cleansing and removing existing duplicate data. This is particular true in the areas of customer master data and product master data. To address this, special software products arose (such as Siperian® which is now owned by Informatica®). Essentially, these products are based on the ETL approach and are then designed to target and service MDM needs, rather than any generic data movement requirement. Informatica® serves this area via the former Siperian® product, which has been deployed for use primarily against customer master data in multiple vertical markets. The open source products have also evolved to support this functionality via MDM specific extensions to their architecture and suites. At this point, it is interesting to note that the open source tools have also evolved over time in very different ways from a variety of starting points. Pentaho™, for example, began primarily as a BI and Analytics platform, but expanded into data movement and integration after that. Talend™ and Apatar™, on the other hand, began life as open source ETL tools and continue to evolve and add features which cross over other areas. Queplix® also supports MDM functionality. Again, its approach is quite different than the other products. Within the ETL suites, MDM components provide out-of-the-box starting points, but then always require custom scripting and coding on a case-by-case basis (based on varying underlying data sources, business rules, and so on). Queplix® provides a user interface for all of these features which allows business users to setup and control MDM processes. This even includes data transformation and implementation of rules via an expression builder type of approach. With full Cloud support and virtualization of data via its metadata layer, the Queplix® suite also supports the concept of "Virtual MDM" which the other products cannot. With the rising popularity of Cloud-based computing and other software hosting and support implementations, this will likely be an important factor for many customers.

Platform Deployment

With respect Cloud computing and other approaches, both existing and expected, we looked across the evaluated products to determine their level of support for the most common configurations in use today and currently evolving. This includes three primary categories: On Premise (or "Behind the Firewall") implementations, SaaS (Software as a Service), and Cloud-based (either Public or Private). In the early days of systems, large proprietary networks provided connectivity to applications within and across businesses via Mainframes and "dumb" terminals. Later, with the advent of "Client-Server" technologies, networks evolved into Minicomputer or Mainframe based servers running a variety of operating systems (such as UNIX, VMS, etc.) with users on PCs that contained "fat clients" (PC-based software containing significant code and functionality). Finally, with the rise of the internet and intranet, the market has leveraged browser-based user interfaces to servers, often with intermediary layers of application or web servers.

Throughout all of this, the On Premise approach has been in existence for the longest period, particularly as businesses developed and deployed their own mission-critical software which was tailored to their needs. All of the products which we evaluated for this study support this option and do it well.

As a standardization and cost savings measure, the SaaS model has become very popular. Companies in non-technology vertical markets have made strategic decisions to move towards the use of packaged applications wherever possible, rather than custom, in-house applications. In addition, with more software firms offering their products on a SaaS basis via subscription, many business have found this to be an attractive alternative to large, up-front license fees. In this space, the ETL tools do not support this approach. Much of the reasoning arises from the prospect of moving large data sets which often reside on site at customers. For an established, up-front license-based vendor such as Informatica®, the transformation to a subscription model can also be a daunting transformation which must be very carefully executed to avoid erosion of existing sales and contracts. The Queplix® suite of products is available via the SaaS model as it was designed with this in mind at the outset. We believe that select partners of Queplix® can provide front-line hosting and/or support as well as Queplix® itself, which provides other options for customers.

Since Cloud computing is the current hot topic in Silicon Valley, two of the open source ETL vendors rapidly stepped forward to add Cloud support to their suites. This is accomplished by one via "snappable connectors" while the other is adding this functionality principally to support one of their related laptop businesses. Either of these approaches puts the control (and onus) on the customer to configure and maintain these Cloud connections. In response to this market need, Queplix® created their own product line which is called "QueCloud™". This service extends beyond simply hosting the software. The company apparently also offers additional data management options as well to assist customers with the ongoing cleansing and harmonization of data between multiple sources.

Connectivity to Data Sources

When examining connectivity to known or possible data sources, we are able to group these as follows :

1. Direct database connectivity (via database specific drivers: JDBC, ODBC, Object-XML, NoSQL, etc.).
2. Application specific API connectors. These are commonly used for mature, enterprise applications such as SAP®, PeopleSoft®, Siebel®, Salesforce®, NetSuite®, etc.
3. Flat files (either fixed length record types or delimited via comma, tab, etc.).
4. Interoperation communication standards such as EDI.

We evaluated connectivity for each of the products based on the most widely used data source technologies currently in use (some of which appear in the bullets above). Table 2 provides detailed commentary by data source regarding the level of support by our pool of products under consideration. For simplicity of presentation, we have grouped the ETL tools together and Queplix® separately, however, any individual ETL product differences are noted when appropriate.

The ETL products do not yet generally provide packaged support for connectivity via Object-XML or NoSQL but Queplix® does. All of the products support every remaining connection method which we analyzed, although they differ by varying degrees in approach. As of the date of this report, Informatica® announced connectivity to NoSQL data types such as Hadoop™. Hadoop™ is fundamentally a file system. It is not a database. It seems far more useful to integrate with a Hadoop™-resident such as Cassandra™.

	Connectivity Type	ETL Products	Queplix
Direct database connectivity (insert/update/delete)	Relational	Yes - all. ETL is designed for relational use.	Yes - Microsoft® SQL Server, MySQL®, Oracle®, IBM-DB2® and Sybase®
	Object - XML	No. Some connectors available which are nominally functional.	Yes - MarkLogic® and others
	NoSQL	No. No connectivity at all.	Yes - Cassandra®, Hbase®, Hive®, and overall support for Hadoop®
Application program interface connectivity (access only through API)	SAP®	Yes - the programmer must have very specific knowledge of the software version and specific API and manage explicit data movement and error handling	Yes - Application Software Blade™ completely hides the proprietary aspects of the vendor API from the business user or programmer - very easy to use. This is a software component.
	Siebel®	Yes - the programmer must have very specific knowledge of the software version and specific API and manage explicit data movement and error handling	Yes - Application Software Blade™ completely hides the proprietary aspects of the vendor API from the business user or programmer - very easy to use. This is a software component.
	PeopleSoft®	Yes - the programmer must have very specific knowledge of the software version and specific API and manage explicit data movement and error handling	Yes - Application Software Blade™ completely hides the proprietary aspects of the vendor API from the business user or programmer - very easy to use. This is a software component.
	Salesforce®	Yes - the programmer must have very specific knowledge of the software version and specific API and manage explicit data movement and error handling. There are some packaged connectors available that convert the object connectivity required for Salesforce® to the relational structure of ETL with associated liability.	Yes - Application Software Blade™ completely hides the proprietary aspects of the vendor API from the business user or programmer - very easy to use. This is a software component.
	NetSuite®	Yes - the programmer must have very specific knowledge of the software version and specific API and manage explicit data movement and error handling. There are some packaged connectors available that convert the object connectivity required for NetSuite® to the relational structure of ETL with associated liability.	Yes - Application Software Blade™ completely hides the proprietary aspects of the vendor API from the business user or programmer - very easy to use. This is a software component.
	Others	Varies - depends on what connector is available. These are very primitive levels of connectivity.	Yes - many other Application Software Blades™ are listed on their website, part of BladeShare™ program
Flatfile		Yes	Yes
Interoperation Standards		Varies	EDI, etc.

Table 2 – Data Source Connectivity Comparison

In this area, as in others, the primary differences between products lie in their approaches. The ETL products are far more "programmer-centric" via the fact that they not only allow for programmatic interaction with the data sources, but in some cases requires it. ETL was designed to build the data warehouse and hence tends to be tightly coupled to relational technology, SQL and tabular data structures. Queplix® is a more modern product, based upon data virtualization technology, and hence is equally well suited to work with relational, object, xml or any other type of data. Discovery is automated and brings data structures to the user's fingertips in a matter of minutes. The Queplix® products abstract all of this via metadata and Software Blade™ approach. In the case of the Queplix® products, data source anomalies are accommodated fully via source specific functionality encapsulated in the blades or via the UI (including logic, expression building).

Synchronization

In the context of our study, the term "synchronization" refers to the overall timing and approach of moving and standardizing data between disparate underlying data sources. The principal approaches to this are Batch-oriented versus Real-time/Incremental (or near real-time, which may mean synching periodically during a business day). All of these products support Batch data synchronization via straightforward approaches. Corporate IT departments frequently incorporate the use of these tools into their existing batch streams via their chosen batch monitoring and control systems (via products such as Autosys®). They also all support Real-time/Incremental updates with some differences in approach. For the ETL products, these updates can be initiated via multiple, regularly scheduled batch jobs (which actually makes the updates "near" real-time). Another method involves enacting strategic triggers on the underlying data sources. However, this may not be possible in practice depending on corporate or IT policies. Queplix® utilizes its own method of non-invasively scanning the data source indexes for date/time stamps which indicate that data has been changed in some way. This is then used to kick off the Real-time update. A core element to the Queplix® offering is "Data Harmonization." This allows users to identify specific records/data elements within all connected data sources as "Records of Truth." Any subsequent change to these records then automatically triggers a harmonization process which replicates the change to all associated data in other underlying sources. Since this is monitored via Queplix® itself, there are no invasive requirements for any of the accessed databases.

Transformation

The movement of data between different systems and technologies often requires some intermediary manipulation or transformation of the data during transition. This may need to occur for various reasons including data structure requirements, formatting requirements, etc. IdealNet compared the products based on their ability to support String manipulation, Mathematical functions, Boolean logic, Fuzzy logic, Statistical logic, Custom transformations, and Summarization.

We are happy to report that each of the products supports all of these requirements. The ETL products provide entry/exit points into which custom transformation code may be implemented. This is primarily done in SQL, although more recent releases now also support routines in languages or script engines such as Java and VBA. Since the Queplix® product contains a complex formula builder with associated UI, all of these transformations are accomplished there. This facilitates the ability for business users to also become engaged in the process where it makes sense.

Data Movement

In the Data Movement category, we looked at three general criteria: (1) bulk data movement (often by large, physical file, (2) data federation (for BI deployment or similar uses), and (3) record- level synchronization. All of the products support bulk data movement, although the ETL products definitely were superior in this aspect. The products all support data federation as well. Although, the ETL software is more time consuming since it often includes programming activity whereas the Queplix® suite leverages its UI on top of metadata to allow for quick data relation specification and rule definition. The same results apply for record-level synchronization as well. In some cases, the ETL tools required SQL programming as part of the process while Queplix® did not.

Test, Development and Operations Environments

Table 3 below details our comparison of the products in terms of their support and capabilities from an environmental approach, including physical environment along with typical corporate and compliance requirements. Here, again, we have grouped the ETL products together and Queplix® based on overall structure and approach.

	ETL Products	Queplix
Target user	Programmer and data center operations personnel	Business user, programmer and data center operations personnel, data steward, data architect (metadata discovery tools)
Dashboard	Programmer workbench with core capability in relational/SQL integration	Object dashboard - visualized objects represent relational, proprietary vendor applications (through their API), object and NoSQL applications.
Operations Console	Varies - basically a sw instance - not a server based product	Full console with alerts for data center environment in on-premise product
Workflow	Some provide limited workflow for data quality	Comprehensive workflow for business logic, data quality and more
Security	Protection for data in-flight between any two applications	Comprehensive support for security model that utilizes existing enterprise security models. LDAP directly supported for users of Queplix products.
Data dictionary - metadata repository	No - schema work is usually external, manual, spreadsheets, etc.	Yes - full data dictionary and metadata repository linked into integration environment.
Shared library of transforms	Varies	Yes
Test & development environment	Yes - not specific to operation	Yes - special permissions enable Test and Development environment
Production environment	Yes - not specific to operation	Yes - production environment. Test environment can be rolled into production environment.
Reporting	Yes - varies	Yes - also includes lineage of data quality, data transforms, problem and exception reporting
Graphic	Varies - some use a graphic driven environment for programmer. All logic still must be explicitly detailed.	No - environment doesn't require programming. Automated dashboard requires configuration and selection - automated alignment eliminates need for graphical programmatic enhancement. Color Automap feature used for semantic analysis.
Disaster recovery	Yes	Yes

Table 3 – Comparison for Test, Development and Operations Environment

Data Modeling

In Table 4, we present our detailed findings in the category of Data Modeling. This encompasses the methods used by the products to locate and present data elements from underlying sources. We refer to the list of tables, records, fields, etc. as "metadata" since these are abstracted from the underlying sources, displayed within the products which we have tested, and then acted upon by users. As you will see, the tools varied widely in this category and not all of them supported every area which we looked at.

	Data Source Type	ETL Products	Queplix
Connection to data sources		Yes - must enter source name, security permissions.	Yes - must enter source name, security permissions.
Discovery of meta-data structure in data sources		No - usually you must be explicitly aware of current data structure.	Yes - complete and full automated discovery of metadata in all potential sources to include relational, object (xml), NoSQL and proprietary applications through API.
Representation of metadata structure in data sources		Yes - usually a tree structure, wire diagram, template or other programmer centric representation	Yes - object dashboard lays out in consistent format with color automat for alignment and semantic discovery
Automatic update of metadata in data sources		No	Yes - automatic even during production operations
Semantic discovery support		No	Yes - color automap, advanced heuristics and application software blades facilitate alignment of relevant fields (objects) for integration
Search of meta-data across multiple sources		No	Yes - fielded search
Direct access to underlying data - in a useful and navigable format - from metadata		Not as part of data modeling. This can be done programmically and explicitly.	Yes - click on a metadata element or browse any catalog directly pursuant to security permissions.
Ability to model all data sources	Relational	Yes - varies. Usually a separate tool required for purchase. ETL is designed for relational tables.	Yes - object dash facilitates all data sources and types.
	Object (XML)	No	Yes - object dash facilitates all data sources and types.
	NoSQL	No	Yes - object dash facilitates all data sources and types.
	Proprietary Applications	No	Yes - object dash facilitates all data sources and types.
Virtual structures in metadata to facilitate mapping		No	Yes - virtual fields, tables (objects) can be created to facilitate BI, data integration mapping and more
Lineage of metadata		No - you must back it up or use separate metadata tool	Yes - virtual fields, tables (objects) can be created to facilitate BI, data integration mapping and more
Metadata export		No - you must back it up or use separate metadata tool	Yes

Table 4 – Comparison regarding Data Modeling Capabilities

Data Quality and Data Governance

When moving and synchronizing various data sources, customers frequently have the need to alter or “correct” certain information to achieve comprehensive data quality. The causes of this are many, but may be precipitated by system error, user error, or general data corruption. For this study, we rated the products based on their support of basic data quality functions as well as more sophisticated, logic-based data governance. Basic data quality activities are supported by all of the products, though the ETL-based software requires a combination of UI and programmatic effort to fully implement. In contrast, Queplix® provides an out-of-the box set of data quality methods which may be selected for use via the UI. Within the ETL products, data governance may be enacted via custom programming procedures, although any level of complexity quickly proves to be extremely difficult in practice. The Queplix® product allows for data governance by including the ability to create a business rule layer which operates before and after data transformations. In keeping with its overall paradigm, a complete user interface is provided to setup and maintain this structure.

Architecture and Standards

Our final area groupings for consideration were Architecture and Standards. Table 5 shows the results for these categories and their respective sub-categories. The support of the products differed as specified and we subdivided categories into certain sub-categories in order to present sufficient detail and provide a clear and fair representation of the results.

Architecture	Category	Sub-category	ETL Products	Queplix®
	Standalone	2 sources	Yes	Yes
		3 or more sources	No	Yes - hub and spoke architecture is designed to scale
	Networked	2 or more integration nodes	No	Yes - in the on-premise product, Virtual Data Manager™, multiple instances can be networked by using a specific Application Software Blade designed for this purpose.
		Virtual MDM	No	Yes, the above blade and 3 additional software modules enable Virtual MDM
		MDM Hub	Yes - must be explicitly and manually programmed	Yes - the architecture can integrate with any legacy or traditional MDM architecture
Standards	SOA		Yes - you must adhere to programming guidelines	Yes
	JDBC		Yes for most	Yes
	ODBC - .NET		Yes for most	Yes
	Web Services		Yes for some	Yes
	Others		Consult individual vendors	Consult individual vendors

Table 5 – Comparison regarding Architecture and Standards

Conclusions

We present this study as a white paper in the hopes that business and technical readers will derive value from the exercise. Having performed and supported a wide variety of data movement and integration projects in various vertical markets over the years, we expected certain results, but were surprised by others. Data integration and movement requirements between two discrete sources can be well served, perhaps with differing costs and implementation models, by all technologies reviewed. Batch-oriented bulk data movement is an area that continues to be best served via the traditional ETL model. Once you consider the need to integrate disparate data types (relational, object, on-premise, cloud, nosql) or more than three sources, we find compelling incentives to use technology other than ETL. This is where the divergence occurs and where the most telling and interesting results reside.

Large, established ETL Vendor

For our study, we selected Informatica® to represent this segment, although we just as easily could have selected IBM®, Oracle®, etc. Informatica® has long been a standard bearer for ETL and, more recently, has expanded its footprint further into the MDM niche area via its acquisition of Siperian®. As the gorilla of the bunch, this offering suffers from similar issues as compared to any vendor of their age, scope, and size. The firm must continue to support a large user base on varying levels of its technology, while attempting to continue to move its products forward simultaneously. While this situation encapsulates recurring revenue streams via long term customer relationships, it can also easily serve to mute the possibility of significant innovation. Based on our examination, we can see that the Informatica® suite of products continues to lead in core ETL specialty areas, but does, in fact lag behind the others within this study in terms of certain key areas of differentiation.

Open Source ETL Vendors

With the relentless acceleration of development platforms and deployment alternatives, open source software has gained legitimacy within even the most staid of corporate environments. Once an open source product has reached a sufficient bandwidth to offer true service levels for support and maintenance, these products become very enticing to customers based on updated technology and methods combined with more flexible price points for acquisition and support. Within our study, we found that all of these products supported core ETL functionality as well as larger vendors. In addition, they each have pursued new and different approaches within their respective product suites which individually offer better performance and features than established vendors. These elements differ by software package and are therefore best evaluated in a case by case basis by customers in response to their specific area of need. The history and ongoing development of each of these open source projects also introduces product tangents which may provide alternative differentiators for prospective use (such as the Pentaho™ BI suite).

New and Different Solution Vendor

Queplix® undoubtedly provided the greatest surprise of all. Even as a small and growing entry, the firm has already demonstrated a truly refreshing and innovative approach to connecting, moving, and harmonizing data across enterprises. The high degree of automation and the attendant ease-of-use, can potentially offer users significant labor and implementation savings over the use of alternate technologies, perhaps on the order of 50% or more. If the integration were to involve three, four, five or more sources, the savings would potentially be much higher. We were impressed with the fact that the Queplix® technology base has been effectively structured to allow the company to rapidly and effectively expand the scope of what can be done with the product. Even some of the most complex and onerous data tasks are handily placed directly under the control of end users via the product's "virtual" architecture coupled with a clean and predictable user interface. Queplix® already has an "A-List" of customers and are very well funded, so we definitely recommend that they be considered as an attractive option.

Author Biography

Christopher C. Biddle

Chief Executive Officer
IdealNet, Inc.

Christopher Biddle has delivered expert Business and Technical services to the Life Sciences Industry (Pharmaceutical, Biotech, and Medical Device manufacturers) and Finance Industry (Securities and Corporate Finance) for 20 years. During this time, Chris founded and has grown two corporations to provide key solutions to clients in the areas of Revenue Management, Finance, and Sales Force Automation. Chris has personally provided consulting services to C level customers at multiple clients including: Deutsche Bank, Bankers Trust New York Corporation, Société Générale, Pfizer, Merck, Johnson & Johnson, Novartis, Sanofi, AstraZeneca, TEVA, Wyeth, and GlaxoSmithKline. In addition, he has also held Partner and Vice President positions with Computer Sciences Corporation (NYSE: CSC), Global Healthcare Exchange (GHX), and iMany, Inc. He has authored numerous papers, articles, and book chapters on business and technology topics. He has also spoken for many years at selected industry conferences across the United States including those sponsored by CBI and IIR. Chris earned his Bachelor's degree in Mathematics from Ursinus College and holds graduate degrees in Management and Finance from Trinity College Dublin and the Wharton School at University of Pennsylvania.

About IdealNet

IdealNet, Inc. provides expert business and technical services to Life Sciences customers and Financial institutions. The firm deploys only senior level resources to solve complex issues for its customers within the specialized areas of revenue management, pricing, mergers and acquisitions, and corporate finance.

IdealNet, Inc.
2933 West Germantown Pike
Building 2, Suite 204
Fairview Village, PA 19409
USA
Sales: 800 618 0836
Fax: 610 666 1006
email: sales@idealnetinc.com
www.idealnetinc.com