# Scholarly Response 2: ETL Tools- Informatica PowerCenter

**Rohan Purohit**

**MSCD 681 Data Warehouse Design**

# College of Computer & Information Sciences

# Regis University

**Scholarly Response 2: ETL Tools- Informatica PowerCenter**

Research a vendor ETL tool, such as Informatica PowerCenter or IBM’s DataStage.Describe the capabilities and functionality of that software. Be sure to include how thetool is used to perform ETL processing and provide your evaluation of the software.What are some capabilities that you would look for to ensure the vendor tool canadequately handle ETL processing and interoperate within your enterprise architecture?Research market trends in the ETL tool market and include this in your research paper.Lastly, provide pros and cons of implementing vendor ETL tools versus building acustom ETL tool*.*

**Introduction**

In modern information centric organizations it is essential to have the capability to integrate data from a wide variety of operational and static sources and stage it to a data warehouse. A data warehouse is implemented as a star schema that facilitates aggregated views of key dimensions and facts for intelligent reporting. A properly implemented data warehouses acts as complex system that enables informed decision making across various enterprise departments. As Ralph Kimball points out in designing the data warehouse, it’s essential to be half MBA and half DBA (Kimball, p.1). Not only should the data warehouse be designed with attention to conformed , cleansed dimensions and consistent and valid facts but it should also be fed clean data and maintained via a thoroughly oiled ETL process. The extraction, transformation and load process improves data quality in the warehouse and selecting an appropriate ETL tool is paramount to ensuring data integrity in the warehouse.

The Informatica PowerCenter Express allows users to design and implement data integration solutions. A data integration developer can use the features of Informatica PowerCenter to extract data from multiple sources, transform the data based on business rules and load the transformed data to target systems. This extract, transform and loading capability highlights the ETL features of Informatica. Informatica also allows users to run profiles that help them to analyze the structure and the content of the source data and subsequently establish data quality. Common data integration tasks and consolidation tasks are made easy by PowerCenter. (Informatica Technical Writers, p.1)

PowerCenter architecture highlights tools, application services and repositories as central components within the Informatica domain. The application clients of Informatica Developer and Informatica Administrator help a user access underlying Informatica functionality. The application clients make requests to the Service Manager that orchestrates the various application services. The application services represent server based functionality offered by key services that the clients make requests to. Repositories are configured as a group of relational databases that store metadata about pertinent objects and processes.

**Analysis and Discussion**

The data integration process is divided into the extract, transform and load phases. First, we create connections to access data from relational databases, web services or social media websites. Then we import metadata to create objects. The next step involves running a profile to analyze the structure of the data we imported. After a profile is obtained we develop mappings to dictate the flow of data between source and target systems. The mappings determine the transformation. We can then define workflows and monitor workflows to tailor business processes.

For example, if we have a collection of flat files that contain individual records from disparate geo locations we can analyze and load this content into a relational database table. Data profiling tools offered by PowerCenter help break down the attributes by quality and precision and consistency. We can then develop mappings between the data elements and monitor the transformation to suit our analysis. The data is then finally loaded into the data warehouse.

**Summary and Conclusion**

Informatica PowerCenter is the market leader in ETL. Currently more than five thousand companies use Informatica PowerCenter to provide data integration services. Training for informatica is very detailed and accessible and support for Informatica has very good customer reviews. Founded in 1993 Informatica has emerged as the dominant force in ETL. Other alternatives include the licensed IBM Data Stage and other open source ETL tools such as Pentaho Data Integration, Talend Open Studio which have thousands of registered users.

Industry surveys carried out in the late 1990’s indicated that all major BI vendors had purchased or developed their own ETL tools during the BI boom (Rizvi, p.2) However these organizations initially assumed that their own ETL process was the right choice, lately this trend has changed indicating that companies are now using fully developed tools by established vendors over in house developed ETL tools or scripts. A case study by the consulting company Passioned in 2009 further pointed that developer productivity would increase by a factor of 3 to 5 times if a proper ETL tool was used. (Rizvi, p.1)

Factors that are important for an ETL are architecture, functionality, usability, reusability, connectivity and interoperability. Outlined by computer scientists these factors have detailed metrics that we can measure for each ETL tool in the market. For architecture we need support for parallel processing, symmetric multiprocessing, clustering, load balancing and have feasibility for grid computing. We should also have multi user support of ETL processes running on multiple machines and have support for common meta-model. (Rizvi, p.3) Functionality dictates that we identify the ETL tool as cleansing based or transformation based, and also identify the support of direct connection to data sources and support of metadata strategy. Usability highlights that the tool should be easy to understand and have good training available for registered users. Reusability notes that we should be able to divide the ETL process into smaller blocks and allow user defined functions to be used in the process flow. Connectivity highlights the support for native connections, metadata reads and the capability to graphically join tables and carry out cleansing, conforming and transformation tasks. Lastly interoperability highlights that the tool should be able to run across various platforms.

In summary the Informatica PowerCenter meets all the success factors required from an ETL tool. It has solid architecture, functionality, usability, reusability and interoperability and support. It is a much more complex and developed offering than an in house developed ETL solution and has lots of support and training material for users. In various scenarios it becomes easier to operate the Informatica ETL tool over natively developed scripts or tools and get the desired results. The product is constantly evolving to support newer technologies like Service Oriented Architecture which is used in modern application integration scenarios. Overall I would recommend Informatica PowerCenter to users who want a professional ETL experience.

**References**

Informatica Technical Writers (2013). Informatica PowerCenter Express Getting Started Guide. Informatica Corporation

Kimball, R. & Ross, M. (2013). The Data Warehouse Toolkit: Third Edition. Indianapolis, IN. Wiley Publishing.

Rizvi, S. & Yousuf, S. (2011). A Comparative Study of ETL Tools. Military College of Signals: National University of Sciences and Technology, Islamabad, Pakistan.