ETL vs ELT

A Subjective View

*Part of the series of the Insource **Commercial Aspects of BI** discussion papers*

Robert J Davenport
June 2008

Contents

Insource is part of the Insource Group, Data Management & Business Intelligence specialists

 Document first published September 2007, revised June 2008

Insource

Insource House, 4 Southern Court, South Street, Reading, Berkshire RG1 4QS

Telephone: **+44 (0) 118 921 1555**
Facsimile: **+44 (0) 118 921 1565**
eMail: **contact@insource.co.uk**
Web: **www.insource.co.uk**

Introduction

Objective

This will provide an in depth understanding and appreciation of the relative strengths and weaknesses of the two different approaches of ETL and ELT in data warehouse design and implementation.

Background

With 80% of a Data Management (DM) or Business Intelligence (BI) project cost being attributable to the design and development of the data warehouse; it is crucial that the right decision is made over the architecture to be employed. Historically, we have had little choice but to use the ETL standard. Emergent technologies are providing toolsets and approaches that challenge this tradition. At least now we have a choice and, as with all choices, it is imperative that, that choice be an informed one.

A wrong decision can be costly and potentially not meet the objectives set out by the project.

Management Summary

In accepting the fact that change is a principle of data warehousing, it is imperative that the approach employed in the design and development of a data warehouse respects, and hopefully, embraces this need. Whilst both ETL and ELT routines can be changed, the process of change in ELT routines involves less cost, time and risk.

Many organizations have recognized the benefits that can be gained from the employment of ELT. In spite of having already deployed warehouse solutions that use ETL, they have embarked upon a process of re-engineering their solutions using ELT through the use of effective tools.

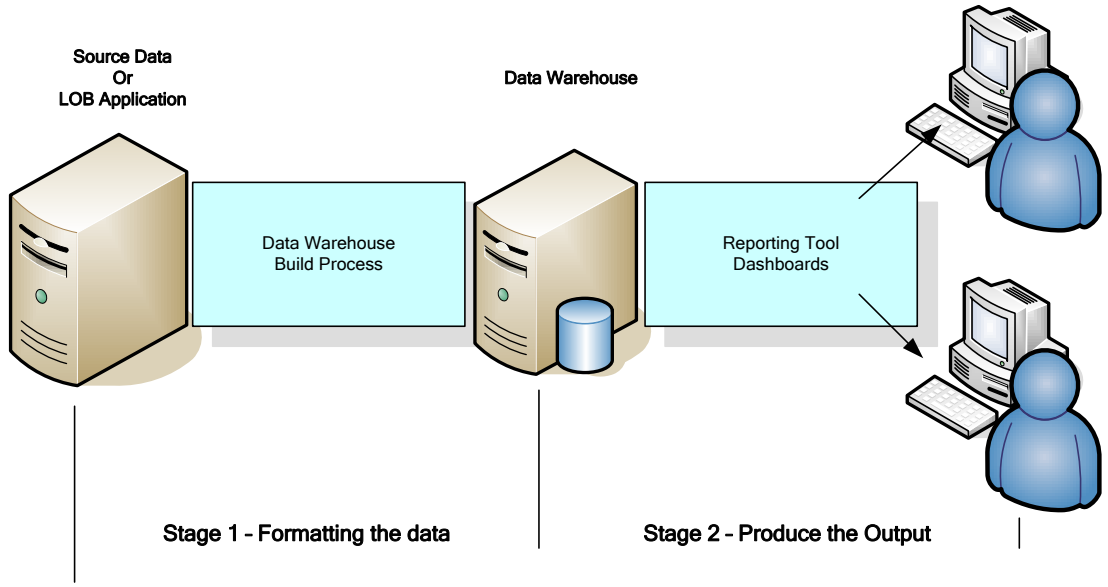
Whilst data warehouse development traditionalists will undoubtedly pour scorn on the ELT approach, the clear and obvious benefits cannot afford to be ignored.

Data Management and Business Intelligence – A Brief Background

As with any discussion document or, for that matter, any debate, it is imperative that definitions are made clear. It is not necessary to go into a lengthy description of what terms will be used in this document, but this next section will make it clear as to the context in which these terms will be used and their meaning.

Data Management and Business Intelligence – is not a technology, or technologies. It is the process by which data is turned into information upon which informed business decisions can be made to effect change. It is therefore a philosophy. For clarity, this document concentrates on the process in the context of the production of a performance management BI solution.

To achieve a DM and BI implementation, two stages are required in the process (excepting any analysis which is part of each).



Format the data – This is the process of producing the data warehouse. It involves getting the data from the source and transforming into an appropriate format from which the output can be derived.

Produce the output – This is the process of taking the transformed data and presenting it to the user in a way that provides information.

The subject of this document is the first of these two stages, “Formatting the data”. To achieve this objective a process must be employed.

Stages in Warehouse Processing

It is generally accepted in informed DM and BI circles that there are three stages in the process of building a data warehouse. However, the order in which these stages are performed can have a fundamental effect on the meaning of these processes and also the overall outcome of the BI solution.

Extract – The process by which data is extracted from the data source

Transform – The transformation of the source data into a format relevant to the solution

Load – The loading of data into the warehouse

Anyone embarking upon a BI strategy believing that it can be defined as a project with a start and end date, is misunderstanding the philosophy and definition of BI. Admittedly, the initial implementation is often where the greatest investment in time and money is borne. However, by the very nature of the definition of BI, once business decisions are taken based on the information presented by the solution, the organization is changed by implication. Using the traditional ETL approach - at best, the BI solution now does not present needed information; portions of the BI solution are potentially redundant, or worse, just plain wrong.

Critically, therefore, decisions taken at the outset of embarking upon a BI solution must respect this fundamental implication and therefore recognize that change is an inherent feature of BI. Therefore, according to this definition, a BI project has no distinct end point and is an ongoing, improving and re-targeting project process.

Employing Tools

There is a distinct difference between what is defined as a technology and a tool. It must be clear about which we are referring to when discussing the subject of data warehousing tools. SQL and Stored Procedures would, for example, be described as technologies. However, as an example, the implementation of Visual Studio included with SQL Server 2005 - namely the Business Intelligence Development Studio - would be described as a tool.

Some third party tools, however, include technologies as part of their packaging. Such tools provide a productivity approach to data warehouse development but, when implemented, use their own technology platform to execute the processes developed.

As database manufacturers recognize the significance of BI, they are increasingly including technologies in the engine that enable more effective development of the data warehouse process, and providing tools to leverage these technologies.

Don’t be seduced by functionality

Their strengths are also their weaknesses. In attempting to provide “all things to all men” functionality, they leave the developer to approach the development process in whichever way they see fit.

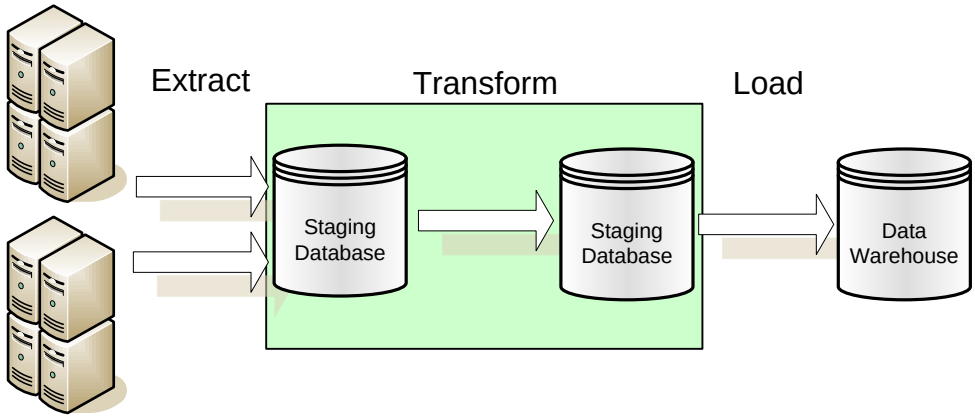
Whilst the end result may well be a working application, the method of implementation will vary wildly. Worse yet, this also results in a mixture of different approaches to the same routines. The practical upshot of which is felt in the difficulty of maintaining the solution. A huge amount of time is wasted in unraveling the methods applied by the original developer. Therefore the only answer is a strict adherence to internal development processes, standards and procedures. This is of course extremely difficult to enforce.

Development process and standards lead to maintainability

The best tools on the market not only provide a flexible platform for development, but also provide a development process that leads the developer through a particular path. This path is therefore consistent across developers and, as a result, deconstructing their work becomes simpler, leading to more maintainable implementations. Remember - change is a fundamental principle of BI and hence data warehousing – and the tools used for data warehousing must be able to accommodate it easily.

ETL – Extract, Transform and Load

This approach to data warehouse development is the traditional and widely accepted approach. The following diagram illustrates each of the individual stages in the process.



Data is “extracted” from the data sources (line of business applications) using a data extraction tool via whatever data connectivity is available. It is then transformed using a series of transformation routines. This transformation process is largely dictated by the data format of the output. Data quality and integrity checking is performed as part of the transformation process, and corrective actions are built into the process. Transformations and integrity checking are performed in the data staging area. Finally, once the data is in the target format, it is then loaded into the data warehouse ready for presentation.

The process is often designed from the end backwards, in that the required output is designed first. In so doing, this informs exactly what data is required from the source. The routines designed and developed to implement the process are written specifically for the purpose of achieving the desired output, and only the data required for the output is included in the extraction process.

In addition, the output design must incorporate all facts and dimensions required to present both the aggregation levels required by the BI solution and any possible future requirements.

Business rules that define how aggregations are achieved and the relationships between the various entities in both the source and target, are designed and therefore coded into the routines that implement the ETL process. This approach leads to tight dependencies in the routines at each stage of the process.

In spite of the fact that there are effectively three stages, the design is often characterized as a monolithic process since the target is predefined and clear.

In addition to the tools provided by the database manufacturer, there are a prolific number of tools available on the market that implement this approach and provide a solid platform, productivity, and flexibility.

Strengths

Development Time

Designing from the output backwards ensures that only data relevant to the solution is extracted and processed, potentially reducing development, extract, and processing overhead; and therefore time.

Targeted data

Due to the targeted nature of the load process, the warehouse contains only data relevant to the presentation.

Administration Overhead

Reduced warehouse content simplifies the security regime implemented and hence the administration overhead.

Tools Availability

The prolific number of tools available that implement ETL provides flexibility of approach and the opportunity to identify a most appropriate tool. The proliferation of tools has lead to a competitive functionality war, which often results in loss of maintainability.

Weaknesses

Flexibility

Targeting only relevant data for output means that any future requirements, that may need data that was not included in the original design, will need to be added to the ETL routines. Due to nature of tight dependency between the routines developed, this often leads to a need for fundamental re-design and development. As a result this increases the time and costs involved.

Hardware

Most third party tools utilize their own engine to implement the ETL process. Regardless of the size of the solution this can necessitate the investment in additional hardware to implement the tool's ETL engine.

Skills Investment

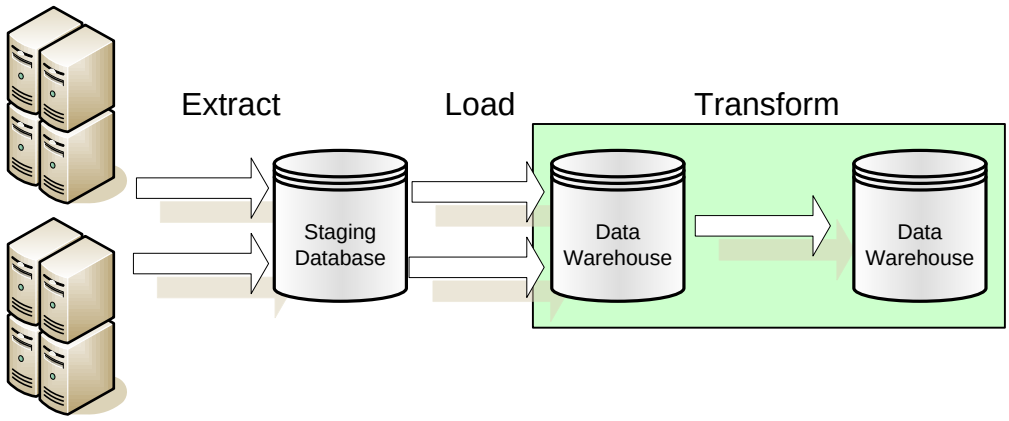
The use of third party tools to implement ETL processes compels the learning of new scripting languages and processes.

Learning Curve

Implementing a third party tool that uses foreign processes and languages results in the learning curve that is implicit in all technologies new to an organization and can often lead to following blind alleys in their use due to lack of experience.

ELT – Extract, Load and Transform

Whilst this approach to the implementation of a warehouse appears on the surface to be similar to ETL, it differs in a number of significant ways. The following diagram illustrates the process.



Data is “extracted” from the data sources (line of business applications) into the “Staging Database” using a data extraction tool via whatever data connectivity is available. Whilst in the staging area, integrity and business rules checks can be applied, and relevant corrections can be made. The source data is then loaded into the warehouse. In effect this provides a validated and cleaned offline copy of the source data in the data warehouse. Once in the warehouse, transformations are performed to re-shape the data into its target output format.

The extract and load process can be isolated from the transformation process. This has a number of benefits.

Isolating the load process from the transformation process removes an inherent dependency between these stages. In addition to including the data necessary for the transformations, the extract and load process can include elements of data that may be required in the future. Indeed, the load process could take the entire source and load it into the warehouse.

Separating the processes enables the project to be broken down into smaller chunks, thus making it more predictable and manageable.

Performing the data integrity checks in the staging area enables a further stage in the process to be isolated and dealt with at the most appropriate point in the process. This approach also helps to ensure that only cleaned and checked data is loaded into the warehouse for transformation.

Isolating the transformations from the load process helps to encourage a more staged approach to the warehouse design and implementation. This embraces the ongoing changing nature of the warehouse build.

Tools are available that inherently implement the ELT process. However, their availability is sparser since this is a more emergent approach to warehouse process design.

Strengths

Project Management

Being able to split the warehouse process into specific and isolated tasks, enables a project to be designed on a smaller task basis, therefore the project can be broken down into manageable chunks.

Flexible & Future Proof

In general, in an ELT implementation all data from the sources are loaded into the warehouse as part of the extract and load process. This, combined with the isolation of the transformation process, means that future requirements can easily be incorporated into the warehouse structure.

Risk minimization

Removing the close interdependencies between each stage of the warehouse build process enables the development process to be isolated, and the individual process design can thus also be isolated. This provides an excellent platform for change, maintenance and management.

Utilize Existing Hardware

In implementing ELT as a warehouse build process, the inherent tools provided with the database engine can be used. Alternatively, the vast majority of the third party ELT tools available employ the use of the database engine's capability and hence the ELT process is run on the same hardware as the database engine underpinning the data warehouse, using the existing hardware deployed.

Utilize Existing Skill sets

By using the functionality provided by the database engine, the existing investments in database skills are re-used to develop the warehouse. No new skills need be learned and the full weight of the experience in developing the engine's technology is utilized, further reducing the cost and risk in the development process.

Weaknesses

Against the Norm

ELT is an emergent approach to data warehouse design and development. Whilst it has proven itself many times over through its abundant use in implementations throughout the world, it does require a change in mentality and design approach against traditional methods. Though some principles remain true to all approaches, ELT is different in many ways. To get the best from an ELT approach requires an open mind.

Tools Availability

Being an emergent technology approach, ELT suffers from a limited availability of tools.

Conclusion

As mentioned earlier, the fundamental nature of BI solutions is one of change and evolution. Employing ETL will undoubtedly limit the ability to embrace this change. At the very best unnecessary cost will be incurred, and at worse significant risk. With the availability of tools that implement ELT extremely effectively, and the clear benefits that can be gained from using ELT over ETL, it's difficult to see why anyone embarking upon the development of a BI solution would use any other approach.

About the author

Rob Davenport is the Products and Technical Services Director for Insource. Being able to architect complex solutions requires an in-depth understanding of DM and BI requirements. Rob has gained this experience over many years as a result of undertaking every aspect of a project from specification, to testing, product development, and training.

Rob is a rare breed individual and a true IT technical business consultant with over 25 years experience in the IT business. This has provided him with a wealth of knowledge and expertise, which has allowed him to apply an equal amount of business and technical consideration to projects in defining their overall solution. He is ultimately responsible for the development program of Insource Data Academy, and WEBdART™, the Insource DM and BI applications with a significant number of successful deployments.

Rob is recognized within the Information Management Industry as an expert on the subject of DM and BI solutions and frequently presents on the subject at seminars hosted by Microsoft.

About Insource

A Microsoft® Gold Partner founded in 1994, Insource provides powerful Data Management, Business Intelligence, Data Warehousing and Reporting solutions. Through the effective combination of our software products, proven implementation processes and knowledge transfer we obtain a clear understanding of your organization's business goals and then provide an end-to-end and/or modular DM and BI solutions.

Data Academy

Insource Data Academy® is a data warehouse builder which leverages the functionality of Microsoft® SQL Server™ 2005 to build robust data warehouses and marts in a fraction of the time it would take using native SQL. This is achieved through the use of intuitive, browser-based screens and data warehousing-specific modules. The resultant SQL data warehouse has full administrative functionality, it can be added to at any time; and because it develops Microsoft SQL native code it has the potential for unlimited further enhancement and customization.

Professional Services

Business Intelligence is more than just technology. From Performance Management to KPI, data analysis, in fact just about any aspect of the organization, it touches on everyone's business life. It involves people, both to bring BI to life and to work within it and around it. It means using the best skills available, taking the knowledge and experience from previous efforts and applying it to your situation, your systems and your procedures.

Insource Professional Services can bring the knowledge and experience from working with some of the most demanding organizations to ensure your BI objectives are met handsomely. From basic help and support to strategic advice and leadership, Insource Professional Services deliver core elements of a project throughout its lifetime - leaving you to concentrate on planning for your new, bright future.

Insource IT Consultancy Ltd

4 Southern Court, South Street, Reading, Berkshire RG1 4QS

Tel: +44 (0) 118 921 1555

Web: www.insource.co.uk

eMail: contact@insource.co.uk

insource
people | technology | results

Microsoft GOLD CERTIFIED Partner

Business Process and Integration
Data Management Solutions
ISV/Software Solutions