# I have given a name to my pain, and it is “ETL”

Bob Strunz, University of Limerick, Ireland

## 1 ABSTRACT

The first stage in the Business Intelligence data flow is the ETL stage, Extract, Transform and Load. This is where data is shipped from the data source into the target data warehouse.

This paper reports preliminary results of an exercise that was intended to assess the potential challenges that might be expected should the institution make the decision to implement an enterprise data-warehouse (EDW).

The results of the experiment were interesting (though not always encouraging) however the process proved to be very worthwhile in demonstrating to senior management some of the more complex aspects of data-storage and management at the institution and also the potential uses that data could be put-to.

The use of Data Profiling on the source database exposed some underlying structural and administrative challenges that will need to be addressed in advance of a full BI implementation.

The implementation of a prototype data-warehouse using Microsoft BI technology(Rainardi, 2008a) demonstrated clearly that the cost of BI need not be a prohibitive factor in an implementation of this type, provided that the institution has the appropriate skills-sets at its disposal.

The exercise of going through the implementation, despite the fact that it was a “whistle-stop” process gave the author a very comprehensive insight into the challenges to be expected and provided the executive of the institution with a range of issues to consider in advance of any implementation decision.

### 1.1 INTRODUCTION

The University of Limerick (UL) is based in the south of Ireland with approximately 13,000 students and 1500 faculty and staff. UL offers a range of programmes up to doctorate and post doctorate levels in the disciplines of Arts, Humanities, Social Sciences, Business, Education, Health Sciences, Science and Engineering.

The University implemented a Student Records System in 1999 which was heavily customised by the vendor at the time to force it to implement existing business processes. This decision is one that the University has had cause to regret in the intervening period because it has rendered the implementation very specific to the institution and consequently it is somewhat expensive to maintain and its business-logic is not always in complete alignment with operational processes on the ground. This divergence leads to inconsistencies in the stored data and complicates the retrieval process considerably when staff are attempting to generate reports.

As currently configured the system provides users with a range of “canned” administrative reports that are perceived to be inflexible and while they are data-rich, they are information-poor. In addition to this, the underlying database (Oracle) can be directly queried using a range of tools to generate more user-specific reports and this practice is quite widespread around the institution.

### 1.2 BUSINESS CASE FOR EDW

The business case for an Enterprise Data Warehouse is a strong one. In the first place, all of the current reporting needs are being met through a range of reports that are derived from the transactional Student Records system. In itself, this is not an optimal approach to operational reporting, the SRS is a typical OLTP system with thousands of tables and the data within it is normalised to an extent that, in the best of worlds, makes accurate data extraction a complex task.

In the less-than-perfect environment of any live OLTP system, data extraction becomes a challenging task even for qualified Oracle professionals.

The second challenge arises from the widespread use of spreadsheets in the institution, this "Spreadmart" culture is rather difficult to eradicate despite its obvious risks and disadvantages (Eckerson, 2003). (Fish, 2014) pointed out that "Spreadmarts are a false step on the BI ladder" they provide the institution with a useful tool but they do not provide it with the stability and flexibility of a properly dimensioned data-warehouse. The author believes that this view, may be challenged to a degree by the latest Microsoft Excel technology which integrates a proper data model into the spreadsheet however while this may be a step on the route to a more structured use of spreadsheets in the organisation it still will not replace the firm yet flexible foundation that a data warehouse should provide.

## 2 PROTOTYPE IMPLEMENTATION

The prototype implementation followed a process recommended by (Kimball, 1996; Rainardi, 2008b) which was straightforward, a single data-mart was implemented on a small portion of the student records system and also an ancillary textual data source.

The process followed was a cut-down version of the recommended approach but it contained all of the stages though they were shortened and were not subject to much user-scrutiny. This exercise was designed to take the implementer through a simulated full deployment but in a very compressed (6-week) timescale and provide senior management with a quick view of what a future system might look like.

It is not intended to cover all of the stages in the process in the context of this paper, the two key implementation stages that are of particular interest are the data feasibility and the ETL implementation which are the areas where we anticipated and indeed faced, challenges. The simple architecture that was implemented is shown in Figure 1.

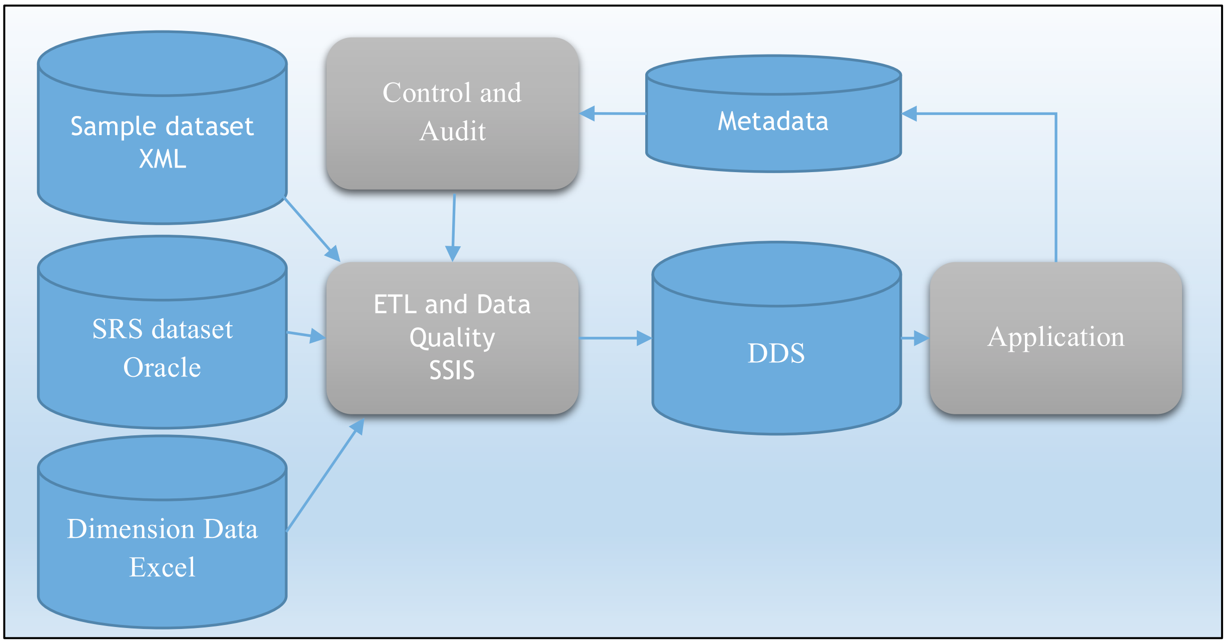

*Figure 1 Prototype Architecture with Combined Staging and DDS ETL*

The DDS is the Dimensional Data Store which is the database that stores the loaded data in a star-schema form, optimised for analysis, where business facts are associated with the dimensions against which the data is to be analysed (e.g. time, gender or programme of study). The SRS data set is the normalised data that is stored in the student records system, it is not optimal for analysis because it is highly normalised.

## 2.1 DATA FEASIBILITY STUDY

The data feasibility study was actually the area where most of the effort was focussed because it was the area where the biggest challenges were anticipated, this proved to be the case and therefore a quite exhaustive field-level study of a small part of the SRS database was conducted. The purpose of this stage is the identification of risks and challenges relating to data.

Data profiling is a powerful tool for quickly evaluating and classifying the structure of data. (Eckerson, 2004) asserts that it has "enormous" benefits when used in the early stages of data warehouse design. There are many profiling tools commercially available but because this was a prototype implementation, the author simply wrote one of his own which was designed to perform a range of simple statistical tests and analyses on selected fields and tables.

These analyses included summarising unique values and the detection of such potential problems as embedded or leading and trailing whitespace in fields, hybrid-coupling (using data range analysis) and referential integrity issues.

The profiling tool was written in a couple of days using existing software libraries and it provided a great deal of insight into how the data was structured and indeed where it was flawed. This process allowed the author to quickly evaluate the probable robustness of any particular ETL approach.

The set of questions answered by the profiling exercise were as follows:

For each field profiled:

1. What is the schema of the data values stored in the field, is it as expected?
2. If the data are used as key fields for other tables is there 100% referential integrity?
3. Is there evidence of Hybrid-Coupling?

It was found that there were significant issues in relation to criteria 1, the data in the fields did not conform to the expected schema, field lengths were often out of specification and supposedly controlled fields often had values that were not legal.

In relation to criteria 2, it was found that there were significant numbers of records which were not stored in a form that was consistent with their expected schema and therefore they could not be reliably joined with some of their associated star-schema dimensions.

These two issues combined make it extremely difficult, if not impossible to write a robust process for the data extraction that can be guaranteed to fail-safe now and into the future. This finding is borne-out by the amount of resources and effort that are directed towards data cleaning and maintenance in the general course of a year at the University. The root causes of a great deal of the issues identified have been traced to an increasing reliance on manual data entry where new academic structures are not supported by the inbuilt business-logic of the SRS.

Criteria 3 Hybrid-Coupling is the use of database fields for the storage of data that does not conform to the meaning of the field as specified in the schema of the database. An example of this would be storage of a Quality Credit Average in a field intended to store Grade Point Averages, if the field stores both types of data, they are indistinguishable from one-another and yet they have different meanings.

Hybrid Coupling was weakly evidenced in some cases however it was strongly observed in others. It is extremely difficult to detect and correct Hybrid Coupling, its effects are very pernicious and it is sadly, a feature of a lot of database systems. Its primary cause is a lack of effective data governance (Griffin, 2010; Otto, 2011) and the only way to counter it is through effective data governance, correcting it after the fact is a difficult task.

The outcome of the data feasibility study was rather discouraging from the perspective of the author, it indicated that there were going to be significant challenges in implementing a reliable ETL process if the cardinal rule of “Zero Tolerance of Error” was to be achieved. This did not however stop the development of an ETL process, it simply limited its scope and outcomes somewhat.

## 2.2 ETL IMPLEMENTATION

The ETL implementation was achieved using SSIS (SQL Server Integration Studio) which provides a quick and easily-learned tool for extracting and transforming data from any data source. The SSIS interface is graphical and the tool supports a wide range of operations on data that would otherwise be difficult to implement using SQL or other technologies (for example pivoting and filtering data).

Figure 2 illustrates an example of an SSIS process to load data from an XML data source (top left) into a series of 3 fact tables (bottom row). The advantage of this approach is that it provides the developer with a very clear and easily maintainable view of how the process works and it also supports a huge range of ready-to-use data processing operations which would otherwise be costly and difficult to develop.

The development of the ETL processes was somewhat circumscribed by the outcomes of the data feasibility study however this can be seen in a positive light as it brought to the surface, issues which would have otherwise been risk factors and these can now be addressed. The primary issue that was surfaced was data Governance and it is from this perspective that all of the challenges will be resolved.

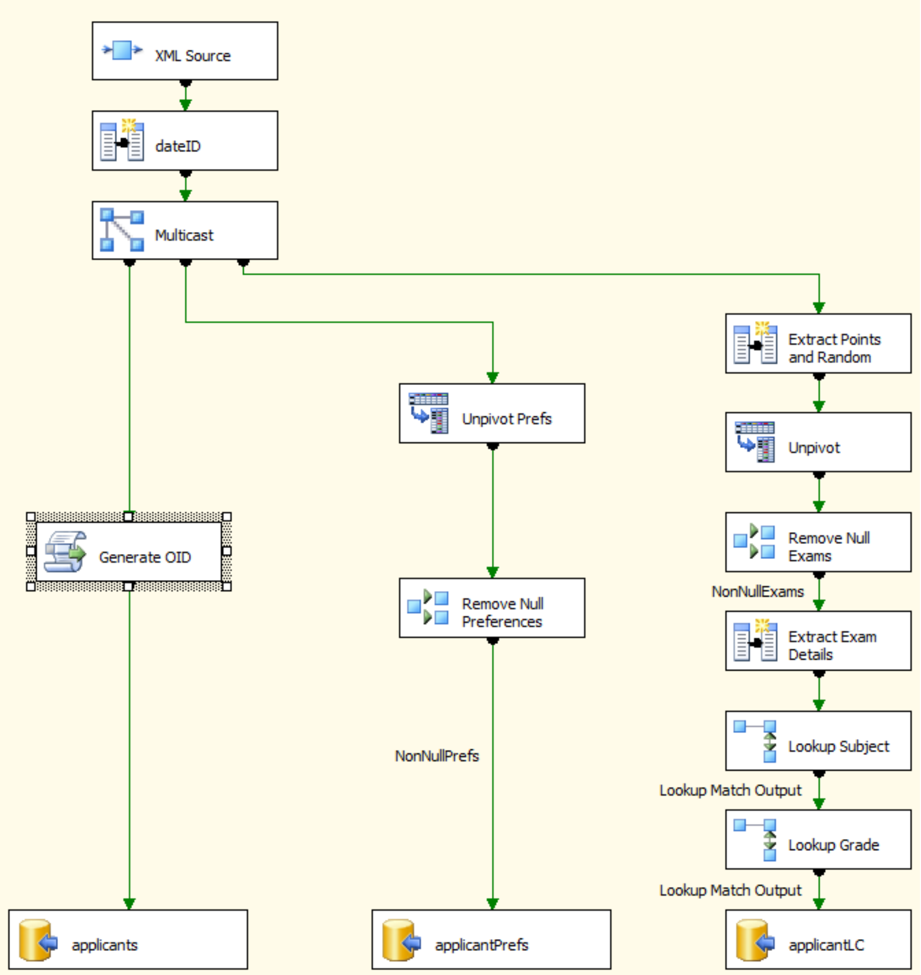

*Figure 2 SSIS DATA Load from XML Data Source*

The use of Microsoft tools for Business Intelligence may seem like a somewhat unusual choice when compared with more common tools such as Oracle, IBM, or SAS but in fact for the University of Limerick, it seems to be a very good fit because it relies on technologies that the institution is already heavily committed to.

The Gartner magic quadrant for BI and Analytics (Gartner, 2013) places Microsoft firmly in the leaders portion of the graph. Microsoft are classed as outright leaders in terms of their ability to execute their vision and are challenging both Oracle and SAS in terms of the completeness of their vision (Schlegel, Sallam, Yuen, & Tapadinhas, 2013). It is very clear that the Microsoft intention and indeed, their action is to develop their Office tools (and specifically Excel) to a completely new level of functionality that can challenge the dominance of traditional players in the BI market. The term "Spreadmart" may actually shift from being a somewhat derisory one to actually being a new paradigm for delivering BI to a mass-market.

# 3 RESULTS

The results of the experiment were extremely valuable to both the author and to the institution. The positive outcomes completely outweighed any negatives.

In the first place, the exercise of actually going through a full development cycle in such a short space of time offers a rich learning experience and affords the researcher and his close colleagues an excellent opportunity to develop their thinking and understanding of the challenges they will face.

In the second place, the conduct of even a limited data-feasibility study helps the candidates to fully appreciate the complexities of the tasks ahead and offers the institution a low-risk opportunity for some in-depth evaluation of its current and future needs.

Finally, the conduct of the exercise allows the development team to identify areas where they may be deficient, in the case of this exercise, it was very clear that the requirements gathering will have to be conducted by an outside consultant because it is too specialised and politically charged internally for people to resolve within a reasonable timescale.

# 4 CONCLUSIONS

It is possible to derive a very high level of value by going through the process that has been described. It is low-cost and low-risk, provided that the institution goes into it with an open mind and with the prior understanding that it will expose issues that may well need strong executive support and a level of investment to resolve.

The approach described presents a minimal level of risk to an institution because it can be accomplished at a very low cost. It is likely that the hardware and software required are already available and the only cost implication relates to the amount of time that the work takes.

One recommendation that is clear from the research is the fact that if an institution made a decision to attempt to duplicate this work it is imperative that the team undertaking the process is blocked-out from other duties. It is the immersive nature of this process that makes it such a rich learning experience.

The final conclusion is that the idea of an EDW is not a hard-sell in most institutions, everybody wants it but the real value of the process is that if affords the institution an opportunity to reflect on how well its business processes are aligned with its systems. The deep insight provided by this is of value, even if the decision to implement an EDW is deferred.

# 5 REFERENCES

Eckerson, W. (2003). The Rise and Fall of Spreadmarts. *Data Management Review, 13*(9), 5.

Eckerson, W. (2004). Data Profiling: A Tool Worth Buying (Really!). [Article]. *DM Review, 14*(6), 28-82.

Fish, O. (2014). *Keynote: Practical Approach To Implementing Business Intelligence in Higher Education*. Paper presented at the EUNIS BI Workshop, AMUE, PARIS, 5-7 March 2014.

Gartner. (2013). 2013 Magic Quadrant for Business Intelligence and Analytics Platforms: Gartner Group.

Griffin, J. (2010). Implementing a Data Governance Initiative. [Article]. *Information Management (1521-2912), 20*(2), 27-28.

Kimball, R. (1996). *The data warehouse toolkit : practical techniques for building dimensional data warehouses*. New York: John Wiley & Sons.

Otto, B. (2011). Data Governance. *Business & Information Systems Engineering, 3*(4), 241-244. doi: 10.1007/s12599-011-0162-8

Rainardi, V. (2008a). *Building a data warehouse with examples in SQL Server*. Berkeley, CA: Apress ; Distributed to the book trade worldwide by Springer-Verlag New York.

Rainardi, V. (2008b). *Building a data warehouse with examples in SQL Server*. Berkeley, CA: Apress ; Springer-Verlag New York.

Schlegel, K., Sallam, R., Yuen, D., & Tapadinhas, J. (2013). Magic Quadrant for Business Intelligence and Analytics Platforms, from http://download.microsoft.com/download/D/D/9/DD94631B-7B68-4F23-870C-C3965FAA222D/2013_gartner_magic_qaudrant_for_bi_and_analytics.pdf

# 6 AUTHORS' BIOGRAPHIES

Bob Strunz is the Technology Advisor to the University of Limerick in Ireland. Bob has a rather mixed background, he did his Diploma and BEng. Degree in Electronics and Computer Engineering and an MEng. Degree in Soil Science and Physics at the University of Limerick, his PhD in Education was done at the University of Hull.

Bob spends his time researching technologies and strategies to make more effective use of technology to improve the quality of the learning experience for students at the University and to provide more and better business intelligence to faculty and staff to enable them to make best use of the increasingly limited resources at our disposal.

Up until Bob completed his PhD he might have been called a “Technical Rationalist” nowadays he sees himself as more of a “Technical Irrationalist”.