Those responsible for data management in your organisation are probably struggling. If you can find an individual or group responsible for everything then they have a lot on their plate. Even if they have managed to write policies and procedures they probably don't have the time or the space to implement and enforce them. So whilst many organisations recognise information as a key asset they are also often unwilling or unable to put the management of data into practice.

Data Management International (DAMA) defines ten Data Management topics in its Book of Knowledge which are: Data Governance; Data Architecture, Analysis & Design; Database Management; Data Security Management; Data Quality Management; Reference & Master Data Management; Data Warehousing & Business Intelligence; Document, Record & Content Management; Metadata Management and Contact Data Management. Ten areas is a lot of ground to cover however, so companies will often understandably start by introducing a number of dedicated initiatives to address specific parts of the problem. Despite this being a positive step, it will often fail as it doesn't take a holistic view in improving the overall management of data.

One such holistic initiative that can work is the creation of a 'Literal Staging Area' or LSA platform. This platform is a literal copy of the business systems created by 1:1 mappings and it is refreshed on a daily basis either by whole dumps or by changing data capture processes. A LSA differs from the concept of an Operational Data Store (or ODS) only in the fact that no compromise is made in the 1:1 nature of the copy that is such an important factor in its maintainability. However, it can also create a further problem in that companies struggle to appreciate how, by adding yet another platform, will help with data management?

The LSA concept started from the Data Warehousing and Business Intelligence perspective with the move from Extract, Transform and Load (ETL) to an Extract, Load and Transform (ELT) strategy where data is extracted from the source system, loaded into the target system and then manipulated for reporting. A well-architected system will isolate each individual source system into its own schema and, by default, create a series of LSAs. Creating this environment immediately reduces the data extraction on operational systems as all downstream systems can query the LSA instead. A further benefit to this approach is the ability to then bring data in from multiple heterogeneous sources that can be used with simple 1:1 mappings. This can also have a further notable effect on the cost of ETL tools where connectors are charged as an additional item. The complex transformations that will come with the population of the data warehouse itself now have a single high-performance homogeneous source from which to get their data.

Once an LSA has been created we can use it as a staging area for the data warehouse. The most obvious secondary use here would be to allow some operational reporting to be done on this system rather than purely at source. If data that is a day old is sufficient for some reporting, and if the data warehouse has finished its downstream processing, then utilising this spare capacity is an obvious choice.

Another use of this data set is for analysis and (re-)design of systems. Often business analysts will require a number of tools along with access to a number of systems. They will also be restricted from using the production systems as a result of performance. Access to a full and complete data set on a homogeneous platform will dramatically reduce analysis time whilst vastly improving the accuracy of results.

Perhaps the least obvious, but largest, return on investment can come from Data Quality Management. This subject is often broken down into two phases, analysis/assessment and cleansing. Whilst cleansing should take place back at source the analysis can be done using the LSA. In fact, it is possible to go much further than a basic assessment or analysis and move the business to adopting continuous monitoring of data quality from which a company can carry out a (very large) number of checks each day in order to track the results over time and identify trends rather than one-off fixes. The scale and benefit from this should not be under-estimated. One current project has added between fifteen and twenty checks to each table in each source system and with an average of around two hundred tables per major source and five major sources this amounts to 15,000 data quality checks daily and consequential trends. All this can easily be managed by a well-designed exception handling process that prioritised trends and reported them to the data quality management team.

All of this seems like it requires a large and complex system but this is not the case. Sizing the system in terms of disk space is an easy calculation as it is the sum of the data space used by the source systems, whilst the mappings (by definition and as described above) are all 1:1. We can also define some other requirements for the type of platform to be successful. Firstly it must be optimised for a very large number of high-performance queries that will allow this workload to be carried out. The solution must also be simple to configure and administer, as the objective is not to add any additional overhead to the systems administration. Finally it must be cost-effective, affordable and scaleable.

Curt Monash (source: www.dbms2.com), an expert in the business intelligence arena, claims that since October 2009 the benchmark figure for systems that can meet this requirement is now around US$20,000 per Terabyte of user data - and in real terms this price point is dropping rather than rising. Monash goes on to suggest that the beauty of systems, such as the Netezza TwinFin - which led the way into this space - is that the number of CPUs in the Massively Parallel Processing (MPP) architecture scales in direct proportion the user disk space.

Taking as an example the system described above where there is massive data quality monitoring requirements across five major sources. The user data from all the source systems amounted to around 3Tb, whilst the data warehouse required around 2Tb and staging areas accounted for another 1Tb of user data space. So, a high performance, simple-to-manage platform for data warehousing with literal staging areas and data quality monitoring & operational reporting can be purchased for around US$120,000.

This type of solution is dramatically changing the challenges for those responsible for data management. Instead of searching for capacity and time to deal with the problems it instead becomes a case of prioritising the activities to make best use of the information available and finding enough business resources to respond to the challenges and issues uncovered by this process.

David Walker
CEO, Data Management and Warehousing
www.datamgmt.com