The Data Game Changer - Part One

Leveraging data projects to spring board your Data Quality initiative

An Experian Data Quality white paper

Experian™
Data Quality

July 2014

Table of contents

About Author

Johny Morris has over 25 years experience in Information Technology spending the last 15 years consulting exclusively on Data Migration assignments for various industries such as Financial services, utilities and telecommunications.

His book “Practical Data Migration” defines the leading Data Migration methodology – PDM. Johny co-founded the specialist data migration consultancy, Iergo Ltd and is on the expert panel of Data Migration Pro.

Author
Johny Morris

Data Quality Specialist

1. Synopsis

When I was asked to write a white paper on why and how a well run Data Migration project can be the spring board to a lasting data quality I assumed that I would be able to deliver it all in a couple of thousand words.

Throughout both papers I will be making reference to Practical Data Migration version 2 (PDMv2) concepts where appropriate.

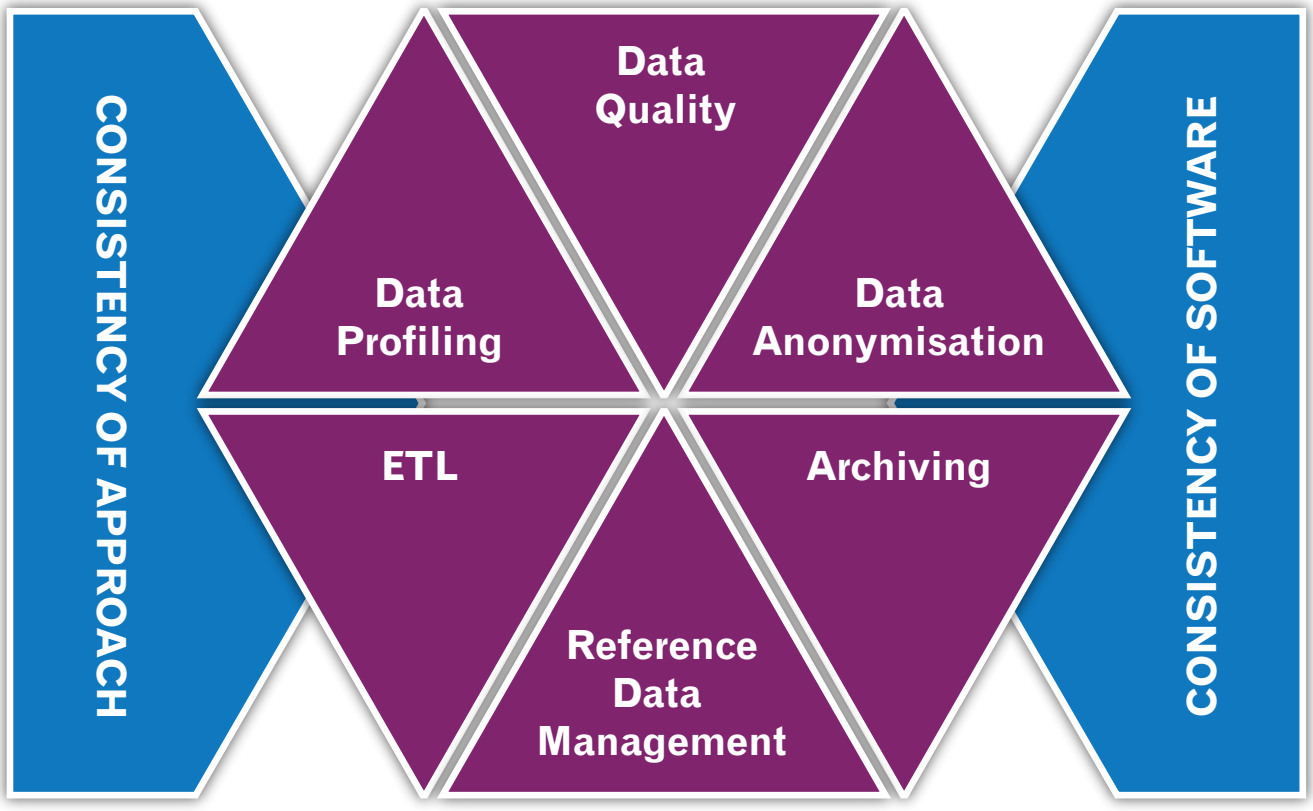
But like a lot of things in life once I had started to write I realised that to explain the general I also needed to explain the particular. In this case that meant I needed to address the software and process side of things as well as why the general pressures that normally get in the way of a data management programme paradoxically become our allies when we are faced with a data migration.

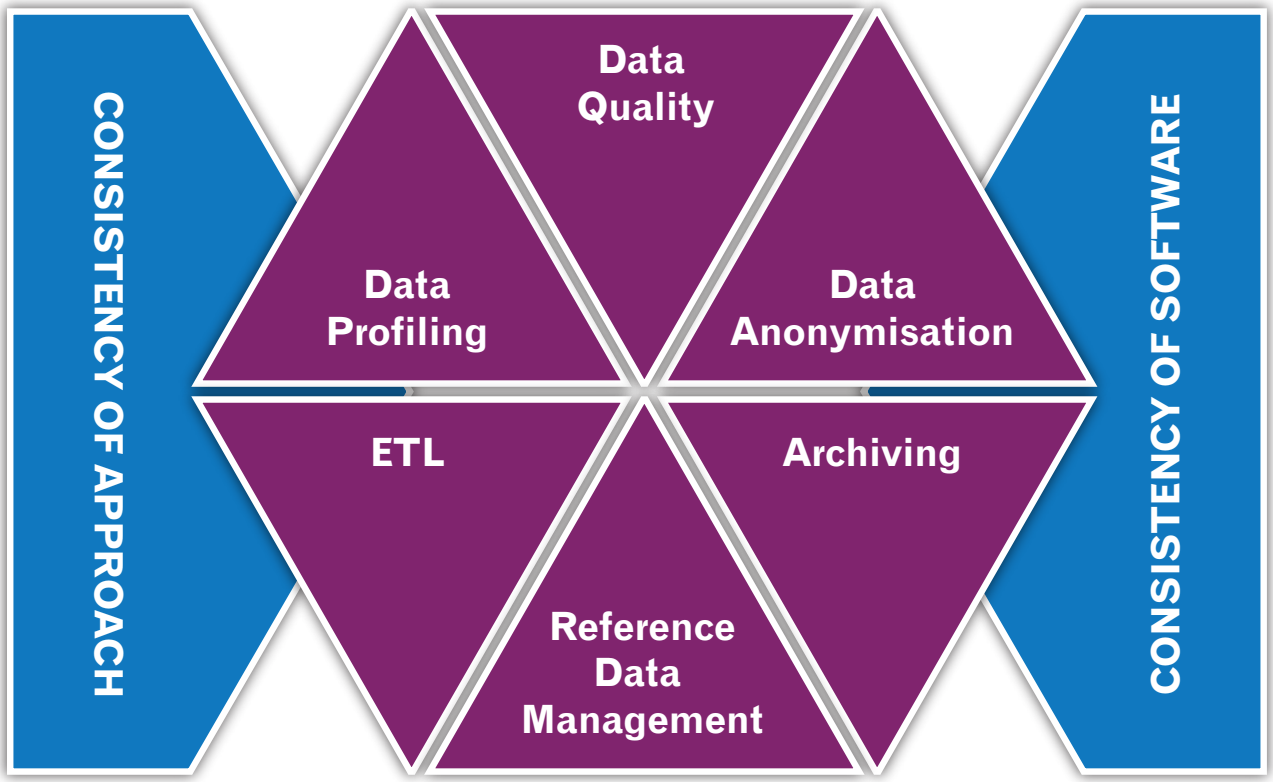
Whilst writing this paper, I found that there were two distinctive sides to the story that would be better represented in two different papers.

This, the first part, covers the stoppers that get in the way of a data management strategy and looks at how data migration is game changing. The second paper looks at software and methodology choices that make it easier to transition from migration to management. Within the second paper we will also look at how to best effect the transition.

Throughout both papers I will be making reference to Practical Data Migration version 2 (PDMv2) concepts where appropriate. This should be no surprise to readers familiar with my other writings – it is the data migration methodology that I authored and contains the features I would like to highlight that can be re-purposed from the project setting of a business change programme to the in-life processes of a data management function but please feel free to substitute any matching best practice that you favour in its place, the arguments within this paper will still stand.

These papers look at 6 dimensions of Data Management and two enablers. I do not pretend to be an expert in data management and from my reading in this area I am aware that there are many views as to the taxonomy of data management so I make no pretence that these are exhaustive. If you prefer, take them as exemplars of some key pillars of your preferred data management framework.





Data Quality

The application of existing business rules to current data to uncover the breaking of those rules in the form of redundancy, synonyms, homonyms, and downright bad data. The existence of these rules can be a consequence of Data Profiling.

Data Anonymisation

The masking of data items to render to protect data subjects or business critical data from deliberate or accidental release into the public domain.

Archiving

The storing of Useful (usually historic) data that is not necessary for the normal functioning of a particular system but for which there is a business need (albeit infrequent).

Reference Data Management

The identification and consolidation of multiple definitions and redundant instance representing the same external object type into a single useful list.

ETL

Extract Transform and Load, the three steps that move data from one data store to another.

Data Profiling

The discovery of the underlying relationships and rules in existing data including the breaking of those rules in the form of redundancy, synonyms, homonyms, and down right bad data.

Consistency of Approach

a single set of steps that are capable of being extended from the Data Migration to the new Business As Usual (BAU) world.

Consistency of Software

a software set that conserves re-usable components and is as suited to the BAU as the Data Migration world.

2. Why is it so hard to get a data management practice established in most enterprises?

When I drop into the conversations of my colleagues in the Data Quality/ Data Management arena the 3 blockers to getting a fully functioning Data Management practice up and running I hear most are:

- **Getting attention, especially of management**
- **The St Augustine Principal**
- **Knowing where to start**

Tackling these individually:

2.1 Challenge 1 - Getting attention

Large organisations are organised along functional lines. There is the production silo, there is the financial spine of the organisation, there is the sales division, marketing and PR etc. etc. These may be further subdivided so that Sales becomes Corporate Sales and Domestic Sales. At the head of each division and sub-division sits a layer of management. And each of these layers can be in favour against or, most likely, indifferent to the Data Management prerogative. However to make headway in a Data Management programme each of the layers needs to be engaged. It's all very well there being a Data Value Chain where the value to the data to consumers is dependent on the quality of data collected by the data gatherers but it is less clear where the value to the data gatherers lies in expending extra time and effort to increase the quality of data beyond the level that they need for their productive activity.

So, for instance, when it comes to choices between more sales calls or making sure the email addresses of telesales successes are captured accurately, neither the person on the phone or the sales director at the top of the tree is under any illusion where their major rewards lie.

I am not here going to get into the use of Key Performance Indicators (KPI) or bonus schemes or quasi disciplinary steps in altering behaviours – this isn't a paper on implementing data management strategies but on using data migration as a stepping stone into getting a data management strategy up and running.

And I am not saying that, in general, most management at any level is antagonistic to better data management. We are all reasonably good corporate citizens and can see the benefit to the whole of altruistic behaviours in our own work practices. Plus most people take a pride in what they do and want to do a good job, so most co-workers are in principal in favour of the proposition that better data will be better for people further down the data value chain and therefore better for the company as a whole. However what stands in the way is what I call the St Augustine effect.

2.2 Challenge 2 – The St Augustine Effect

As we all know St Augustine of Hippo was an early Christian church leader, bishop of Hippo in North Africa and author of a number of philosophical and theological works. He is the patron saint of brewers amongst other things on account of his pleasure seeking life as a young man before he converted to Christianity at the age of 33. In his autobiographical and theological work "Confessions" he famously has his young self declare "make me good lord, but not yet" as he plunges on with more sins of the flesh.

This could be the cri de coeur of many of our colleagues when faced with the prospect of actually doing something about the data we know to be wrong and the processes we know could be tightened up to prevent bad data leaking into the food chain. We would really like to do something about it. We are all in favour. But just not this week/month/quarter. We have another more pressing issue to resolve. There is this month's sales conference to organise, this quarter end's figures to reconcile, this year's major sales period to manage. And so it goes – the tin can of bad data gets kicked down the road until we forget about how good it could be with better data and learn to live within the limitations.

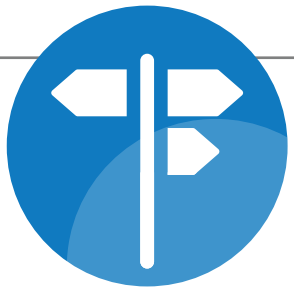
But even where there is a will to change and a budget provided and a large slice of senior executive sponsorship on top there is still the issue of where to start.

2.3 Challenge 3 – Where to Start

In this white paper I am looking at 6 dimensions of good data governance but they are all interrelated. Do we start with say a single view of customer (Reference Data Management)? This implies that we tidy up all the definitions and records of Customers wherever they are held (Data Profiling and Quality), bringing them to a suitable single repository before replication across multiple platforms (ETL). Each of these choices limits other choices we could have made and the whole piece looks overly complicated and therefore doomed before we start.

An alternative strategy is to start small. Maybe we cleanup one area only? Perhaps we can remove those overhanging work orders that have never been completed in the system although we know they must have been completed in reality. Possibly we could look at removing the blockers to getting our invoices processed in time? This kind of tidy up is attractive. Start small, show results and then move on to bigger and better things. But scaling is still a problem. As soon as you move out of the local, within silo, activity you encounter Challenge 1.

There is also the problem of sponsorship. Making the internal sell to get approval (and budget) for something big enough to make a difference is hard. To justify the trouble you really need to show a success that justified the effort – especially the effort on the part of the executive to give your proposal air time over the hundred and one others that there are out there.



So given these blockers to success how does leveraging the reality of a data migration help you to overcome them?

Let's start with my definition of what a data migration is. Within PDMv2 a data migration is:

> **"The selection, preparation, extraction, transformation and permanent movement of appropriate data that is of the right quality, to the right place at the right time and the decommissioning of legacy data stores"**

Each part of this definition is significant and pertinent to the underlying lessons in this white paper:

Selection. We often have a choice of where to get our data from. There are the legacy corporate systems with overlapping data sets, there are the local departmental systems which may have more accurate data, there are the spreadsheets that really run most companies and may have the most accurate or least accurate information. In this welter of possibilities where do we turn? The answer of course is in our Profiling. We will look in detail later in the paper at how we might use software to help us accomplish this but clearly in a well managed data migration we will be doing really deep and really thorough data profiling. We need to be able to read across heterogeneous sources to find redundancies, gaps and omissions. One data set may tell us that Customer A likes to be addressed as Al not Alistair. Another data set may tell us that Customer A prefers to be contacted on his mobile not his office number. We need to bring these things together.

There will be datasets that are supposed to link but to what degree is this actually true? Do we have orphan records anywhere? All of which leads us onto **Preparation.** We know we will find records that need to be corrected. How are we going to achieve this? Well this comes in the preparation of the data. This can cover anything from enriching data from third party sources to manual correcting of data in the legacy. It can also be accomplished in the **Transformation** of data during the flow from source to target systems. We will have to transform data anyway because the target and the source are never the same. If they were there would be little reason to change systems. This of course implies the use of an Extract Transform and Load (ETL) tool. Again as we shall see later when we look at software specifics, these can be functions embodied in a single tool or they can be distributed over a number of tools.

Moving data of the **right quality** is interesting. Please note that I do not say 100% perfect quality data. In my extensive experience no company wants, needs or will pay for perfect quality data. In the world of real projects there are constraints of time and budget which are often tied in to the business case that preclude pursuing data quality beyond a certain point. What is needed is rigorous prioritisation. And for that we need to know all the issues we are going to face as soon as possible. All of which leads us back to profiling.

Data quality is also interesting because we do not work in static worlds where things are not changing. As we progress through a project whatever it was that caused the bad data in the first place may still be going on. Data is still being corrupted. So again what is needed is a constant monitoring of the quality of our legacy data. How are we progressing to our targets for data quality? Will we hit them in time to go live? Or are we sliding backwards with some data items that were correct being re-corrupted?

To monitor this appropriately, on enterprise level projects we need enterprise strength tools and the processes to go with them.

So we need profiling tools, data quality tools and ETL tools.

3.1 But what about Reference Data Management, Anonymisation and Archiving?

Obviously it is frequently the case that we need to perform Reference **Data Management** as part of migrations. This is all bound up in data quality and data loading. Whether it is a product hierarchy, the relationship of currencies to regions or a de-duplicated customer list, in data migrations we always need to keep track of reference data lists. Not only do we need to keep track of them we need to generate them in the first place.

Now this does not mean that we need to have a fully fledged Master Data Management (MDM) package in place. Mostly our lists will be used as either input to the data quality activity to qualify which data items do not pass the data quality threshold or as parts of our data preparation transformations during ETL. We will not (usually) be replicating these lists back into the legacy systems. We are performing most of MDM but (usually) without the replication.

When it comes to **Anonymisation** it seems there is an increasing trend to expect the data migration work stream to provide data not just for the target at the end of the project but also for a number of intermediary activities along the way. It may be data for testing, training or filling the sandpit database but, given the strength of data protection legislation we will need to anonymise it. This is often more complicated than it looks because, for the data to be useful , the same anonymised versions of linking data items will have to be created. Having tools that will identify and allow this to happen, thus preserving referential integrity, is imperative.

All of which brings us finally to **Archiving.** We know that we will not be able to bring over all the historical data the old systems contain for a number of reasons.

Firstly the older the data the more preparation it needs. This may be a rule of thumb but it holds good pretty much everywhere. Over time business structures change, regulatory rules change, validation rules change. Over time there will have been a series of system crashes and bugs (fixed now maybe). All these things leave their mark on the data in the legacy data store. Drilling back through the data is like drilling back through time.

Secondly there are some simple problems of the target load programs and how they work. It is certainly the case with many COTS (Computer Off The Shelf) systems (Microsoft Dynamics, SAP to an extent) that the load programs run through the application middleware. Each item gets loaded as a part of an appropriate transaction. So to load historic, closed, purchase orders a new purchase order needs to be created in exactly the same way it was in the source with part delivered purchase order lines etc. that are subsequently filled , then it needs to be completed. All of this will be done using the transactions that would have been used if the invoice had been managed manually in the target. This is both highly risky in terms of preparing sufficient data accurately and submitting it in exactly the right sequence and highly costly in terms of time. Completing a transaction at a time is slow. Way slower than the good old days of writing data directly into the database.

However there is always data that is needed (albeit rarely and often not accessed at the same speed as data routinely used during face to face interactions with customers). I am thinking here, for instance of the 7 years of tax records that might be asked for one day or the whole life data retention requirement placed on social service and medical systems. None of these need to be available real

time/online but when requested must be forthcoming in an agreed time frame (although the statutory time limit for some of these can be counted in weeks) when requested. They must also have an unimpeachable data lineage (data lineage is the trace of how a data item was captured transformed and stored).

So we routinely have a need for an Archive solution.

3.2 The Compelling Event

Which brings us to the biggest advantage we have over the Data Management guys – we are in the business of "decommissioning Legacy Data Stores". This gives us our compelling event. We are switching production systems off. Ok, so sometimes this is a logical switch off. The system may still be in use either for other parts of the same organisation or for another enterprise in de-mergers, but from the point of view of the business with which we are engaged, the legacy system will not exist when we have finished out work.

The benefit this gives us is that when we are talking to senior managers in parallel silos we can get them to focus on our requirements by pointing out that the systems in question are nearly at the end of their lives. Indeed I always recommend that on the very first meeting with managers outside of the technical areas (and even there in many cases) do not use the words data migration at all. Make it clear that you are here to switch System X off. Give your audience a clear date when it will happen. Tell them that six weeks before switching the old system off you will be back, asking for them to sign off on their permission for you to turn this system off (a decommissioning certificate in my language). To follow the St Augustine metaphor, this confirmation of mortality always provides the healthy nudge your business colleagues need to commit to working on the preparation, data quality, reference data management and archiving activities you need. And to start that activity today.

3.3 Keeping the momentum

It is often said that it is easier to maintain a Mercedes than it is to build one in the first place and this is true of data management practices. Once they have proved their worth and have momentum they are easier to keep going than they are to get up and running in the first place but to achieve this three things need to be in place:

- Suitable software that can carry over re-useable components from the data migration
- Suitable methods and approaches that can transition from the high pressure of project life to the more enervating atmosphere of business as usual
- And of course appropriate levels of sponsorship so blockers in the way can be removed

So the opportunity to springboard into the data management process you always wanted is there for you if you make the right choices within your data migration.

In "The Data Game Changer Part 2" we will look at the software, process and crucial influencing strategies that will underpin this transition and discuss how to make the leap.

About Experian Data Quality

Experian Data Quality has built up exceptional market coverage assisting customers with their unique data quality challenges.

We provide a comprehensive toolkit for data quality projects combining our market leading software with a vast scope of reference data assets and services. Our mission is to put our customers in a position to make the right decisions from accurate and reliable data. The size and scope of data management projects varies considerably but the common factor in all ventures is unlocking operational efficiency and improving customer engagement. We see the potential of data. Whether it's in enabling ambulances to be sent to the exact location of an emergency or attributing charitable donations to the people who need it the most - data accuracy makes all the difference to service provision.

For more information please contact us:
T (61) 3 8699 0100 | E info@au.experian.com | W edq.com/au

© Experian, 2015. All rights reserved
The word "EXPERIAN" and the graphical device are trade marks of Experian and/or its associated companies and may be registered in the EU, USA and other countries. The graphical device is a registered Community design in the EU.