

**Data Quality:
The Key to Managing the Successful Data Warehouse Project
By Peter B. Giblett**

One of the common mistakes made during the planning of a data warehousing project is to assume that data quality will be addressed during testing. To make such an assumption will lead to either delays in delivery, or at the extreme total non-delivery of the project. To leave checking of data quality till testing will leave no time to address the issues raised and to put into place corrective measures then still insure that the project gets delivered on time.

It is essential to start investigating the organisation's data quality whilst initiating the project. Early warning of quality issues will empower the project to take, or initiate, corrective measures.

It may seem strange to start an investigation that is essentially technical in nature so early in a project lifecycle. But there are very good reasons for commencing this at such an early stage. Few, if any, projects start from the proposition "we need a data warehouse". The majority of projects start with a desire to measure the performance some part of the organisation. The business community often globalises this idea before involving the IT Department. Even before the project commences the business has some idea of what is to be modelled and where the data will come from.

The purpose of the project is to clarify the aims and to define a business requirement, decide elements that are in/out of scope and deploy processes to ensure successful delivery.

One of the major stumbling blocks to successful delivery of the Data Warehousing project is always data quality. Almost every organisation has a data quality nightmare within their operational systems right now. (See **examples 1 and 2**). Before you jump up and deny this I shall add that these problems have little or no impact on either the operational systems or the departments that use them. These problems will have a major impact on the data warehouse.

The business impact of poor data quality can include:

- Loss of credibility of the system
- Business community dissatisfaction
- Delays in deployment
- Extra project time to reconcile data
- Cancelled project
- Lost corporate revenue opportunities
- Compliance problems

Conversely the benefits of high data quality can include:

- Greater confidence in analytical systems
- Increased business community satisfaction

- Less time reconciling data
- A single version of the truth
- Reduced costs
- Ability to identify corporate revenue opportunities

There is one basic truth about any Data Warehouse: if the information in it is not proven to be correct then the business community will not accept it. The technical functions may deliver all the data required, but this is no guarantee of ultimate acceptance. It is the functioning of the data quality initiative that will guarantee the successful business intelligence delivery.

What is Data Quality in the context of a Data Warehouse?

Data quality is driven by the need the expectations of the knowledge worker. It encompasses the need for data that is devoid of errors, but also meets business demands for:

- **Completeness:** All the data sent arrives at its destination.
- **Accuracy:** Can be verified against existing information delivery mechanisms.
- **Validity:** Rules exist to ensure that all data is processed, and that the business is aware of all exceptions.
- **Consistency:** Extract and transformation rules enable a continuing and accurate process.
- **Accessible:** All data is accessible understandable and usable.
- **Integrity:** Supporting databases continue to ensure that entities, attributes and relationships are maintained consistently.

Complete, accurate, valid and consistent data are the keys to source data quality. Data needs to be accessible in an open format and provided in a timely fashion. This process needs to be repeated every time data is loaded into the warehouse. Ultimately once the system is in production it will be the role of the Warehouse Administrator to ensure that the latest incremental data load is consistent with these rules.

How does the project meet these criteria?

Some organisations have already taken the step of creating a Corporate Data Standards and Quality unit within either the business community or IT. If such a unit exists the project must seek out this body and work with it, agreeing a strategy for the project and the business. The existence of a Data Quality unit does not remove from the project responsibility to analyse data quality, it may provide a shortcut or simplify the job.

Most organisations have no such unit; it is therefore necessary to take on the role of the Corporate Data Quality unit within the scope of the warehouse project.

Diagram 1 shows the major steps to be taken in managing the Data Warehouse project. It shows that Data Quality is not only an integral part of the project, but will remain a sustained and ongoing activity.

I have previously stated that the Data Quality Initiative should start almost as soon as the project is conceived. This initiative should be staffed by at least one person from the business community, who has both in-depth knowledge of operational systems and has the vision to pursue the goals of the project. Tracking down the causes of poor quality data obviously requires technical knowledge; therefore the initiative should also be staffed by at least one IT person. The scale of the project will determine the size of the Data Quality team.

This team should be empowered to investigate any systems that are likely to supply the warehouse and where possible take copies of live data for investigation (within a test or QA server). UNDER NO CIRCUMSTANCES are they allowed to assess, or impede the operation of live systems.

What is the role of the Data Quality Team?

Diagram 1 shows that this team is required to be integrally involved with the whole project from the business requirements definition through to the user acceptance testing. The team's role will include:

1. Review existing business processes.
2. Review existing technical and data architecture.
3. Assess data quality.
4. Make specific recommendations for cleaning of data.
5. Set-up a continuous monitoring process.
6. Either recommend corrective measures are implemented within the source system or are brought into the scope of the project.

Review of business processes, will lead to a better understanding of how the business works with its systems. The review of the technical and data architecture will enable the team to recognise the key technical components and how they link together. These reviews are also required when defining business and technical requirements

During the early stages of the project, for example when the Business Requirement is being defined they need to decompose business ideas, understand the systems from which data will be sourced and assess the data quality. The aim is to identify both the strengths and the potential pitfalls. This will act to provide early warning of probable issues e.g. daily totals not balancing, or essential date and numeric data being stored in free-text fields.

The earlier problems are identified the sooner specific recommendations can be made to ensure that data is correctly cleansed. Recommendations must either identify corrections to be made to source systems or ensure that corrective steps are designed into the ETL process from the outset.

Continuous monitoring is necessary to ensure that when fixes are made to source systems other unforeseen errors do not occur as the Warehouse is built.

What will be the result?

A project that will deliver a fully functional information set, with results that are proven and believable.

Checking for data quality within a Data Warehousing context demands more than mere testing. It demands a pro-active approach, with a team that is looking for problems before they become a problem. To identify a problem at the time when the first data load occurs is to doom the project to failure. A Data Warehouse project rests not on the fancy graphical reports that it produces, but on whether the figures contained within it are right.

In an OLTP system data quality will form part of the testing process. In theory the whole process defined within this article should not be necessary, the OLTP system has been tested and the company has successfully relied on it for years.

The truth is that changes have been made over the years that have eroded the integrity of the OLTP system. In some cases problems have been fixed, yet old data remains unfixed. In other cases additional data has been transferred into the system on its journey through corporate life. Business rules that apply to the applications are not always implemented so rigorously on data transfers.

The fact remains that all data required to feed the Data Warehouse is prone to error. One of the rules of IT is that all imported data should be treated as suspect until it is proven valid. This rule is an essential part of the quality control process applied to the warehouse. It is not merely a rule for the testing phase of the project. It continues throughout the lifetime of the warehouse. It is the duty of the Warehouse Administrator to continue to monitor the quality of data within the Warehouse.

Longevity is the key to information quality in the data warehouse. Solving problems early will ensure success.

Example 1

A large airline has a many systems relating to many business areas, Bookings, Departure Control, Flight Operations, Cabin Services, Baggage Control and Catering. Each of these systems reports a different number of passengers on-board each flight journey.

Whilst the departure control system is the preferred source of passenger data it is inconsistently operated in different airports through the world. Additionally some flights, in remote locations, are checked in using another company's departure systems. Data is eventually sent to the airline the quality of the data is poor and includes repetitions. E.g. 4 Mr Smith's of 3 Acacia Avenue - Is this data repetition or is it a family of four entered under the same name by a hurried departure agent?

Example 2

A manufacturing company, with plants around the world, recorded all faults through a central help-line. Faults are allocated to a Resolving Agency, which could be either an internal workman, or an external maintenance company. Progress on all faults was recorded in the help-desk system. Within the help-desk system important information stored in a free format note field:

- such as the date/time of actions taken (in an inconsistent format),
- the time spent on resolving the fault and
- occasionally the engineer's name.

