Experian™
Data Quality

Data Quality Management Software

Experian Pandora

Experian Whitepaper

Contents

Who should read this document?
You should read this document if you are interested in improving the profitability and reducing the operational and compliance risks of your organisation through business-focussed management of your data. Whether your organisation uses terms such as Data Quality, Data Governance, Compliance, or Data Stewardship, the common process which supports all these activities can be described as Data Quality Management.

Document synopsis
The document describes the common tasks that make up a Data Quality Management process and the software tool capabilities that can enable and support those tasks. After outlining the drawbacks commonly encountered when setting up such processes, we detail how Experian Pandora can provide the necessary capabilities and highlight some unique features and benefits enjoyed by users of Pandora.

99% of enterprises now have a Data Quality Management Strategy, yet 94% of large companies admit that Data Quality issues are still having a negative impact on costs, revenues and business risks*. - **The Data Advantage, Experian, 2013**

1. Data Quality is impacting your business

Data is at the heart of every organisation, and Data Quality is a key ingredient of performance, profitability and business risk.

Since business decisions based on incorrect, missing or inconsistent data are not likely to be optimal, it is not surprising that 99% of enterprises now have a Data Quality Management Strategy. Unfortunately, 94% of them admit that Data Quality issues are still having a negative impact on costs, revenues and business risks*.

Research** shows that 40% of the anticipated value of all business initiatives is never achieved. Poor data quality in both the planning and execution phases of these initiatives is a primary cause. Poor data quality also effects operational efficiency, risk mitigation and agility by compromising the decisions made in each of these areas:

- Poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits
- Data quality effects overall labor productivity by as much as a 20%

Some problems, such as duplicate or incorrectly addressed mail, are not business-critical, but are costly. By decreasing returned mail by 10%, a healthcare company realized $400,000 in savings over three years***.

Unfortunately, many quality data costs are much higher but less visible. Data Quality issues affect all areas of business, from customer service and regulatory compliance, to sales and marketing productivity and financial accuracy:

- Issues such as un-invoiced goods, payment for goods never received, incorrect product deliveries, overstocking, stranded assets and mis-reported financial results are usually the result of poor Data Quality.
- Delays caused by Data Quality issues can even lead to companies being late-to-market with the resultant opportunity cost.
- Data quality issues can lead to fines being imposed by regulatory bodies, an example being when the rules surrounding the storage of personal information are not adhered to.

Meanwhile, increasing volume, velocity and variety of data are conspiring with increasing time pressures from business to make Data Quality targets ever more difficult to attain.

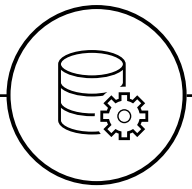
Although most organisations have a Data Quality strategy and people are actively trying to address data quality issues, they are rarely equipping themselves with the tools necessary to succeed. Two thirds of companies today are using manual techniques in an attempt to manage Data Quality*. Some employ software tools, but these are either rudimentary and disjointed or technically complex and expensive to adopt.

Unsurprisingly, results are disappointing. For example, the Libor Rate discrepancies for which several banks were fined hundreds of millions in 2012/2013 could have been uncovered and dealt with internally if a tool had been set up to systematically check for inconsistencies in the data between the relevant systems.

In this paper we will list the capabilities required to manage Data Quality across the Enterprise and highlight common constraints of existing approaches. We will then describe the unique capabilities that Experian Pandora can bring to Data Quality Management in any organisation.

> "The speed at which we could get projects signed off and approved was unreal. Our first insurance customer was absolutely amazed with the productivity. Very, very powerful."
>
> **Charles Thomas, CGI**

* "The Data Advantage", Experian, 2013
** "Measuring the Business Value of Data Quality", Gartner, T. Friedman & M. Smith, 2013
*** "The importance of data quality in producing savings", Healthcare Finance News, 2007



2. Data Quality Management software requirements

In order to combat the costs, lost revenue and business risks caused by data quality, organisations need to establish a data quality management process to understand, improve and monitor data quality. This process should be part of the everyday life of the organisation and in order to be efficient should be supported by software tools.

2.1 Basic capabilities of a DQ process

Firstly, a Data Quality process needs to provide teams with the ability to analyse data and to transform or correct it.

2.1.1 Data analysis

Data analysis should be performed against real data, read as-is from production systems to ensure that we have a true picture of what's actually there. This should be carried out using a Data Profiling tool, which:

- provides a statistical analysis of data, some technical (completeness, uniqueness) and some rule-based (product codes require a particular format)
- allows investigation of particular cases

Data should whenever possible be analysed in its entirety as sampling techniques are not 100% reliable.

Data quality assessment is often based on comparison of values which reside in different computer systems. The process should be sophisticated enough to allow this cross-system analysis. Examples are "reconciliation", "consistency" and "integrity".

2.1.2 Data correction

The process should allow team members to fix the data. The fixing process, whether manual or automated, can involve transformation, standardisation, enrichment, consolidation, selection and de-duplication. This can be accomplished by:

- Transforming the data directly in the source systems, and for small volumes this is a good approach. The process simply needs to identify which records need modified and the necessary values
- Transforming the data off-line and feeding it back to the originating system via the relevant IT team
- Developing and validating the correction rules and providing a specification for the IT team to implement

2.2 Capabilities of a good DQ process

The most effective Data Quality processes contain much more than the basic capabilities described in the proceeding section of this document.

- The process must focus on providing the organisation with the best return on investment in the shortest time
- The right people need to be allowed to bring their expertise to bear
- The process should attempt to fix the root cause as well as the bad data
- The whole process must be manageable

"You can deploy Pandora very quickly. Business users like it because it's easy to get to grips with and it is straightforward and easy to use."

Richard O'Neill,
Data Quality Manager, Financial Times

Carrying out workshops and discussing data on-screen is **extremely productive** and can **cut out days of email communication**.

2.2.1 Fast time to result

Organisations should not have to wait weeks or months to fully understand the implications of a data issue, or how to tackle it. The sooner a plan of action is decided, the sooner the losses can be stemmed or the benefits achieved. The ability to provide a result quickly is therefore vital.

2.2.2 Business relevant & accessible

A good DQ process associates business-defined relevance to its findings. Financial transactions can give a simple way of weighting the importance of issues, however some "weighting" of results requires investigation and calculation. However it is done, the process must associate a value or weighting to each issue to allow prioritisation and to justify doing anything at all about an issue. There's no point in spending £10,000 to fix an issue that's costing the company £100.

2.2.3 Able to prioritise tasks

The scale of the problem and its cost should already have been established by earlier investigations.

Now, by analysing the bad data in detail, it should be possible to estimate the cost to fix it. For example, it can be the case that 95% of the data can be quickly corrected with some automatic rules, leaving 5% to be corrected manually. The more accurate the understanding of the potential solution, the better the estimates, so an environment which allows the data fixes to be quickly "tried out" and evaluated allows better-informed decisions.

Armed with these details the business can take informed, business-relevant, objective decisions about if/how to fix Data Quality issues.

2.2.4 Collaborative

Neither IT nor the business people have all the answers concerning data quality so one of the most important capabilities of any process is enabling team members to collaborate in a single, easy-to-use environment. The process should allow team members to concentrate on data quality, not on whether they can write clever scripts.

Carrying out joint workshops, looking at, manipulating and discussing the real data on-screen is extremely productive and can cut out days of email communication or discussions based on descriptions and expectations.

2.2.5 Summary & Detail reports

Management statistics about Data Quality are necessary, and their presentation should be flexible to allow "by system", "by business area", "by quality Dimension" and so on. Such statistics need to be provided over-time to indicate progress, and therefore return on investment.

It is also necessary to provide the associated record-level detail so that issues can be investigated by team members and solutions proposed.

2.2.6 Proactive data analysis

Data analysis has traditionally been about validating (or not) expectations about the data. There are three types of data issue:

- **The bad data you know about**

  Some of these issues are well understood, but many are "urban legend" and may not even exist. As you investigate you will understand the scale of the problems, and the knock-on effects they are having elsewhere.

- **The unexpected data that's not actually wrong**

  Examples of this include data fields being re-used, or unexpected values (a company which only delivers to its domestic market but which has customer addresses in another country).

- **The bad data you didn't know about!**

  This is the worst, because it is unexpected, and may already be having a very significant impact on the performance of your organisation.

A good Data Quality Management process will proactively propose information about the data and allow analysts to find issues they weren't necessarily looking for. A field which is 99.9% complete should be suspicious, as should a 7 character product code when all other product codes have 8 characters.

2.2.7 Issue Management

Once uncovered, issues need to be documented, backed up with evidence and examples, and be managed. It should be possible to navigate from the issue to the actual data easily in order to carry out further investigation. Example data should follow standard Data Governance procedures, meaning it should probably not be sent as screen shots attached to an email.

2.2.8 Able to find root cause

Fixing the data is only one half of the job. If at all possible, the root cause of the issue should be fixed too, otherwise you will find yourself with more bad data tomorrow. Bad data is the result of a flawed business process, so detailed analysis of the bad data, looking for trends and common attributes, is required to home in and find the flaw.

2.2.9 Reusable

Rules used to validate, standardise, transform and match data should be available for re-use by other Data Quality Management processes and ideally by applications across the enterprise.

Reference data gathered/compiled for use by Data Quality rules should be re-usable and maintainable.

2.2.10 Data Assets inventory

Knowledge gained about enterprise data, its location, quality, ownership, associated rules, etc. should be stored for re-use by colleagues.

People rarely find issues they are not looking for, so **finding unexpected issues requires a great deal of luck**.

3. "Best Intentions" are not enough

Despite all their hard work and best intentions, results are slow to obtain, and both IT and business people find the Data Quality Management process extremely frustrating.

3.1 Typical approaches

The typical approaches to Data Quality are as follows:

Traditional Data Quality tools concentrate heavily on cleansing name and address data and are renowned for being technically complex and difficult to integrate with other software solutions. Furthermore, most of these tools were designed decades ago, so have significant technology, volume and performance limitations.

Many organisations rely on writing manual scripts to answer Data Quality questions from their business colleagues and to correct data. The very nature of this type of investigation, where each response leads to further questions, makes this approach slow and error-prone. The scripts used to correct data are based on the poor results of the investigation, so their logic is often incomplete or inaccurate. Microsoft Excel is a very popular Data Quality tool among business people because it is easy to obtain, and it provides a quick, flexible result when compared to what IT departments can provide with DQ tools or scripting. Of course Excel is used only on small datasets, remains "single-user", and is very labour intensive. The functionality is very limited unless the user wants to start programming, and consistent re-use is impossible.

3.2 Typical issues

Existing approaches to Data Quality are usually restricted to fixing known issues with name and address data, but they do not provide an end-to-end framework for managing all data quality. Existing approaches have the following drawbacks:

3.2.1 Slow to result

Whether because of technical complexity in use, difficult installation, or because they only address a small part of the overall process, existing approaches are slow to provide results. Often, IT and business blame each other for lack of results, but whatever the reasons, the long time to result is there for all to see. And because solutions do not allow easy re-use, each new task has to start from scratch.

3.2.2 Not business-relevant

Providing record counts and percentage Data Quality scores is better than providing no information, but business people always struggle to see the relevance of such information. Existing approaches do not put a relative weighting or monetary value on issues, so they remain statistics rather than business-relevant facts. This lack of business relevance could explain why business people are keen to have quality data but remain suspicious or even dismissive of the Data Quality processes implemented.

3.3 Typical causes

3.3.1 Difficult to collaborate

Since traditional approaches rely on many technical components, tasks and specialist skills, team members find themselves doing different tasks using different tools for each task. Collaboration is often no more sophisticated than exchanging notes at team meetings; experience shows that this does not really constitute collaboration.

3.3.2 Only validating expectations

Existing approaches to analysis, transformation and monitoring are based on whether the data conforms to expectations. People rarely find issues they are not looking for, so finding unexpected issues requires a great deal of luck.

If you don't examine the whole hay stack, how do you know you've found all the needles?

3.3.3 Reporting is summary-level but not actionable detail

Existing approaches often provide data quality statistics in the form of a percentage. While this may be interesting to management, it is not "actionable" and unfortunately the necessary detail is not usually accessible or usable when trying to find root cause or determine fixes.

3.3.4 Separated from issue management

Fortunate teams have issue management software with which to work, however these are not well adapted to Data Quality issues. While they do provide workflow capabilities they do not allow easy association of "evidence" and are frustratingly separate from the data itself. Evidence can be data from multiple systems and can include millions of records. Often teams send emails with screen shots of data, making issue management difficult and representing a data confidentiality/governance issue in itself.

3.3.5 Single-system, based on samples

Usual approaches address data in one system at a time. The ability to investigate, compare and combine data from multiple systems is usually overlooked. With the increasing amount of data being managed by organisations, approaches to Data Quality are based on data samples for all but the smallest data sets. Basing your understanding of the contents of a dataset on choosing

a "statistically representative" sample relies on a perfect understanding of the data, which is what we are trying to achieve in the first place... The approach of using samples is clearly contradictory to the objective of data analysis. If you don't examine the whole hay stack, how do you know you've found all the needles?

3.3.6 Unable to allow ad-hoc investigation

Most approaches to Data Quality investigation involve writing and running scripts which produce record counts or show particular values or records in isolation. None provide a no-limits interactive capability to do ad-hoc investigation on full data volumes across all systems. Having to wait a long time for each answer causes analysis fatigue and analysts naturally limit their investigations to simple cases.

3.3.7 Little help with root-cause

The ability to carry out detailed ad-hoc analysis on "bad data" is extremely rare and always constrained by volumes. Yet this is essential for finding and fixing root causes in a timely manner.

3.3.8 No inventory

Few existing approaches include the ability to create an inventory of data, with standardised definitions which associate similar data across systems. When valuable information of this sort is gathered it is not made available for re-use on other projects.

4. Experian Pandora for Data Quality

The Experian Pandora software product allows organisations to carry out enterprise wide Data Quality Management and Data Governance and brings a step-change in Data Management productivity thanks to its easy, interactive functionality. With Experian Pandora, Data Quality Management is quick, collaborative and transparent. It enables business-focussed Data Quality activities leading

to improved data consistency and validity, and driving improved business processes, higher profitability and lower costs. Experian Pandora can be installed and providing valuable data insight within minutes, and is used by technical and non-technical staff, facilitating adoption across

the organisation. Experian Pandora enables organisations to analyse and investigate data, assess and improve its quality, and manage this process over time, independent of whether the information concerns customers, products, finance, sales or other business areas. Pandora provides all the functionality required to make Data Quality Management easy. In the rest of this paper we will describe how Pandora provides all the capabilities required for a good data Quality Management process, followed by a description of the significant extra functionality which we believe makes it unique.

4.1 A Data Management process with Experian Pandora

4.1.1 Proactive Data Quality Analytics

Data analysis underpins the entire Data Management process, and there is no better way to analyse data than with Experian Pandora. This high performance analytical engine allows organisations to build a complete understanding of their data easily and quickly without depending on specialist technical skills or third-party products.

Experian Pandora provides analysis which can be categorised as Data Profiling and Data Discovery.

- Data Profiling is universally acknowledged by data management professionals as an essential activity for all data-dependent projects. Its objective is to provide understanding of data.

- Data Discovery concerns the uncovering and investigation of relationships between data items.

This understanding feeds the activities of feasibility or risk assessment, project scoping, data integration mapping design, data quality improvement planning and testing. Philip Howard of Bloor Research wrote "In the context of our research into the market for Profiling and Discovery tools Pandora has the highest score of any product for its architecture and the same is true of analysis."

A Pandora user simply has to select a table or file from the list proposed by Pandora, for it to be completely analysed using a series of sophisticated, pre-defined rules. By analysing 100% of the data across any number of systems, Pandora proactively provides the most complete and accurate information possible:

- Technical profile (completeness, uniqueness, formats, largest value, ...)

- Rule-based profile (product codes require a particular format, column B is 20% of column A)

Pandora automatically provides statistics about the data, discovers relationships, implicit data rules and inconsistencies, and allows investigation of all cases, carrying out "reconciliation", "consistency" and "integrity" checks when required.

In fact, subject matter experts build their own analytical system for Data Quality to underpin all subsequent data-dependent activities.

> "In the context of our research into the market for Profiling and Discovery tools, Pandora has the highest score of any product for its architecture and the same is true of analysis.
>
> **Philip Howard, Bloor Research**

Pandora provides an out of the box Data Quality report with a summary for each column. This report is interactive, allowing instant navigation to any underlying statistic and associated data record. Pandora automatically derives and calculates over 200 different pieces of information about every column, enabling analysts to quickly spot and investigate issues:

- Frequency distribution of column values, formats and phonetic patterns
- Unique value, format and phonetic counts
- Actual and most prevalent data types
- Smallest, largest, least common, most common values and formats
- Shortest, average and longest value length
- Numeric scale & precision
- Sum, average, variance & standard deviation
- Null and blank counts

Rule results are stored to provide time-based directional KPIs.

Pandora provides a secure, collaborative environment where colleagues can chat, exchange reports, share data and re-use rules.

Team members are given only the data access and capabilities necessary, with security being applied instantly and dynamically, e.g. colleagues could view the same credit card table but some may see scrambled versions of the credit card numbers.

Users document and manage issues in Pandora, giving each a priority and category, and assigning them to a team member. Pandora stores each contribution to each issue, providing a clear indication of progress and automatically building an issue management audit trail for the project manager.

4.1.2 Collaborative monitoring and reporting

Data Quality rules provide individual scores which can be consolidated by system, by dimension, by product line etc. Over-time trends are automatically provided. All summary statistics can be expanded to show the actual underlying records, allowing teams to investigate interactively and take appropriate actions.

Analysts use Pandora to interactively analyse bad data in detail in order to uncover common attributes and patterns of behaviour. Incorrect data is usually evidence of a broken business process, so by analysing this evidence the process can be identified. Armed with the calculated value of the issue (from earlier in the process) and the probable cost to fix the data an appropriate, cost-justified solution can be chosen.

4.1.3 Data correction

The same interface used to analyse the data is used to fix it. Pandora provides interactive functionality to transform, standardise, correct, enrich, consolidate and de-duplicate data. Along the way it can join, filter, group, union, sort and perform lookups.

Instead of writing rules into a spreadsheet or document, analysts and subject matter experts collaborate interactively to design the required processes using Pandora, quickly and easily building data prototypes in the same software tool that they used to perform their data analysis.

Pandora applies each transformation to the data in real time (filter, transform, standardise, join, sort, group, sum, count, etc.) and the analyst can immediately see the resulting data on-screen choosing to validate or correct it. This prototyping activity continues until the data has been fixed, ready to be sent back to the relevant systems.

This approach of validating a result as opposed to a collection of hand-written rules provides results extremely quickly and represents a step-change in Data Management productivity. Pandora's graphical transformation builder is entirely point-and-click and does not require programming, scripting or any other technical skill.

Analysts use interactive analysis and data prototyping to analyse bad data in detail and quickly try out correction rules.

The effort required to carry out potential fixes can be evaluated and compared with the measurement of the problems established earlier in the process to allow objective, business-focused prioritisation.

4.1.4 Reusable rules and inventory

Rules developed to validate or transform data can be saved and re-used. Reference data, for validation, standardisation, identification, categorisation, etc. can be created, saved and re-used.

Pandora tracks rule and reference data usage automatically, providing instant impact analysis for changes. The Experian Pandora Business Glossary allows Stewards and Subject Matter Experts

to define business terms and associate them with physical data assets.

4.2 Pandora extras

Customers state that Experian Pandora is quick and easy to use, allowing short "time to value" and rapid return on investment. Pandora brings a step-change to Data Management productivity.

The following functionality is what we believe sets Experian Pandora apart from other solutions.

4.2.1 Business relevance - monetisation

Analysts can attribute a weighting or monetary amount to issues by associating rules with a "measure" – a numerical value which is either present in the data or can be derived using values in the data. This provides business relevant financial KPIs.

This allows issues to be compared and prioritised. Data Quality activities can be justified with respect to a value and/or business priority. While this is not unheard of in Data Management it is very rare, and extremely technical to set up.

Pandora puts this capability into the hands of non-technical team members, and does not require any programming or scripting. Pandora also has the benefit of allowing data from multiple systems to contribute to the rules and value calculations.

4.2.2 Easy, cross system collaboration

Pandora provides a multi-user environment where functionality and access to data can be defined for each user. This means that all team members use the same interface and can share data and findings, chat, contribute to discussions on issues, and re-use rules, reference data and definitions. The security controls are simple to manage and are applied dynamically, so when two colleagues view the same table one may see a scrambled version of sensitive data in a column.

The Pandora software is able to read data from files, spread-sheets and database systems, providing a "global" view of data without the need to build a central staging are or repository. This allows cross-system integrity analysis, reconciliation and validation without requiring masses of technical infrastructure and large technical teams to support the Data Quality Management team.

The software helps analysts to spot the issues they were not looking for.

The Pandora Outliers report is a unique user-customisable analysis which compares the data statistics derived by Pandora with a normal distribution curve and highlights statistically unusual data – it find issues that no-one was even looking for.

To assist the Analyst Pandora also analyses data with respect to a "standard distribution", and highlights values which are "unusually" long, short, frequent or infrequent, or have an unusual "format".

Pandora also indicates if there are missing values in a numerical sequence, which could indicate a record has been deleted.

To assist the Analyst Pandora also analyses data with respect to a "standard distribution", and highlights values which are "unusually" long, short, frequent or infrequent, or have an unusual "format". Pandora also indicates if there are missing values in a numerical sequence, which could indicate a record has been deleted. The interactive nature of Pandora allows ad-hoc investigation, interrogation and Data Quality analytics across all data and systems, comparing and combining data regardless of its provenance. Uniquely, this does not require technical expertise and does not need any other software; Pandora provides its own self-defining, self-optimising repository.

4.2.4 One-stop with low TCO

Experian Pandora is a single, easy to use software solution providing all the functionality required for a Data Quality Management process.

Since Pandora does not require any other software products and its maintenance tasks are automatic, there are no hidden costs. When compared to other approaches or solutions which require database licenses, development tasks and regular maintenance operations, the total cost of owning Pandora is extremely competitive.

4.2.5 Work with full data volumes

Pandora is able to cope with hundreds of millions, or billions of records per table or file, so its results for analysis, transformation and reporting are 100% reliable. Not only is the initial analysis of the data very quick, the subsequent interactive analysis, search and investigation is interactive and instantaneous, with all questions being answered directly by Pandora without the need to go back and query the original source. Time to result is very low when compared to every other approach.

4.2.6 Flexible, reusable, callable rules

All rules in Pandora, whether validation or transformation, standard or user-specified, can be re-used by colleagues developing other processes within Pandora as well as from outside Pandora.

Rules are available to other enterprise applications via the Pandora API, allowing a consistent Data Quality process to be established across the enterprise.

5. Conclusion

The Experian Pandora software supports an end-to-end Data Quality Management process, driving a range of benefits to the business through prioritised Data Quality tasks:

- More profit thanks to better business decisions
- Less customer churn thanks to improved customer management
- Operational cost savings thanks to improved business processes
- Earlier, proactive actions to resolve DQ issues cost less than reacting to problems
- Easier and quicker Compliance reporting
- Optimised ROI due to business-focussed prioritisation of Data Quality tasks
- Easier, more objective Data Quality business cases
- Easier, more efficient team communication and tracking
- Lower risk, faster delivery of data-related projects
- Better use of business experts' time

6. Experian Software

Experian Software is a leading Data Management software vendor.

The Experian Pandora software product combines business-minded usability, unrivalled performance, and the richest selection of capabilities to accelerate organisations towards Data Management maturity.

Experian Pandora is used by customers in a diverse range of industries across the world to carry out activities such as Data Quality Management, Data Governance and Data Migration.

Experian Data Quality
George West House,
2-3 Clapham Common Northside,
London, SW4 0QL.
T 0800 197 7920 | E dataquality@experian.com | W www.edq.com/uk

Intelligent interactions.
Every time.

© Experian, 2013. 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.

Experian Ltd is authorised and regulated by the Financial Conduct Authority. Experian Ltd is registered in England and Wales under company registration number 653331. Registered office address: Landmark House, Experian Way,