Data Management through the value chain

Stay in control of your reporting flow

Paul Rothwell, Ruurd van der Ham, Marinka Voorhout and Daniël Foudraine

Organizations experience an increasing demand for high quality data due to a rise in analysis techniques and the availability of data, as well as increasingly demanding regulations and legislation. However, this demand for quality is not limited to the data residing in the source systems. It has become clear that control over data quality should cover the entire flow of data: from source to report. This gives organizations the opportunity to achieve true data driven reporting and decision making, but also brings along several challenges that need to be overcome.

P.A. Rothwell
is a partner at KPMG Financial Management.
rothwell.paul@kpmg.nl

R.F. van der Ham
is a senior consultant at KPMG Enterprise Data Management.
vanderham.ruurd@kpmg.nl

M.A. Voorhout
is a senior manager at KPMG Enterprise Data Management.
voorhout.marinka@kpmg.nl

D.A. Foudraine
is a consultant at KPMG Enterprise Data Management.
foudraine.daniel@kpmg.nl

Introduction

The importance of good data quality is eminent within various sectors and industries ([Jonk12]). It is evident that having more and more access to data from various sources proliferates the importance of high data quality. The need for high data quality is also stimulated by increasing possibilities for analysis and reporting purposes. It is however less clear that this increasing demand for high data quality simultaneously increases the complexity of data. For example, high data quality is moreover demanded to improve organizational performance, support growth, competitive advantage and comply to the growing demands of data driven regulations.

This seems like a contradiction. Having a strong focus on data should gradually resolve quality and maintenance issues throughout an organization. For some organizations this is true for their source systems (i.e. master data management). More and more companies realize that the importance of data quality is not limited to their source data. It extends to the flow of data within their reporting chain (the so-called data flow). This flow requires usage of consistent data definitions and data quality criteria, from source to report.

A focus on this flow of data is relatively new and can be seen throughout different sectors. Internal and external users of reports are wondering if their control information is timely and fact-based, supported by good data quality. Various sectors such as financial services, medical devices, telecommunication, pharmaceutical, the public sector and consumer markets are subject to stricter regulations regarding data collection and usage ([Voor16]). Supervisory authorities are shifting their traditional reporting (output) based monitoring towards data driven supervision, requiring proven data quality consistently used within reporting chains (input and throughput).

The quality of data from source to report

The focus on data quality within reporting chains is seen across industries. Due to progressive regulation as a reaction to the financial crisis, organizing data quality is maturing most rapidly within the Financial Services sector. Various legislation is impacting the data management:

1. BCBS #239 regarding the collection, disclosure and usage of data right down to the quality of internal decision making;

2. an extensive and common set of standardized reporting templates complete with built-in data validation rules (Data Point Model);
3. extensive data-sets at transactional data level (AnaCredit) opening the door to true data-driven reporting (Banking Integrated Reporting Directive).

These requirements all have one thing in common: data quality is paramount and subject to review by external parties such as supervisors. Portfolio decision making can suffer material consequences if data is incorrectly defined and classified. Examples of this are in the Financial Services sector where risk weighting depends on the correct channeling of data into the appropriate portfolios. Failure to do so can severely impact capital and liquidity positions and put the bank or insurance company and its customers at risk. All this strongly increases the need for improved data quality throughout the reporting chain.

It is not only the Financial Services sector, however, that is working on improving the data flows in their reporting chains. In general, the more complex the organization becomes, the more risk it has of not having adequate steering information. For example, a large global energy company with local plants, regional offices and several locations for their head office wants to get a grip on their data flow. Asking themselves relevant questions such as: which data is used within local plants to derive reports, what data quality criteria is applied and which data transformations take place along its flow? In other words: how can insights into the quality and consistency of the data flow be derived?

Within the retail industry there is also a focus on legislation which is driving the need for data flows. Although not (yet) as extensive as financial regulations, the EU regulation on the provision of food information to consumers combines EU rules on general food labeling and nutrition labelling into one piece of legislation.[1] The new regulation makes nutrition labeling mandatory, and instructs food manufacturers to provide information on the energy value and six nutrients. This means for retailers that they need to have access to the data which was created earlier in the manufacturing supply chain, i.e. by the supplier. Given the fact that retailers have a multitude of suppliers a so-called data pool (governed by GS1)[2] has been set up that functions as a storage facility for both the food suppliers (delivering input into the data pool) and retailers (using the data pool to be able to inform their customers with specific data about, for instance, allergies). Thus setting up a complete data flow where suppliers, retailers and consumers deliver and use consistent, correct and timely data.

Main challenges in managing data flows

Whether organizations are driven by increasing regulations within their sector or because there is acknowledgement that the environment requires fast and flexible insights and fact-based decision taking, a growing number of organizations are transforming to a data-driven organization. Those organizations have already found that improving their existing data management and usage activities is usually experienced as complex. This is caused by amongst others unclear ownership, limited understanding of data (quality) and the tendency to be convinced that data management is an IT department responsibility rather than a business responsibility. However within this complexity, capturing and improving a data flow (or reporting flow) has turned out to have its own distinctive set of challenges.

Understanding data flows

- Unfamiliarity with compliance at the required granular data level means that organizations have the tendency to back away from it, especially if data is transported and transformed. Particularly when transformations are complex, it can require specialist effort to determine which data elements refer to one another;
- Organizations have difficulties in distinguishing a process flow from a reporting flow, assuming that existing process flows can function as a data flow overview;

> *In general, the more complex the organization gets, the more risk it has of not having adequate control information*

[1] https://www.food.gov.uk/enforcement/regulation/fir/labelling

[2] www.gs1.com

Particularly when transformations are complex, it can require specialist effort to determine which data elements refer to one another

- Lacking overview of the Key Data Elements (KDEs) that are used for different reporting purposes within systems, departments, processes and End User Computing solutions (e.g. Excel) within the reporting flow – especially if employees are only familiar with their (silo-based) tasks and cannot oversee the complete or even partial flow, nor the materiality of these KDEs;
- The complexity of tracking data increases in companies with complex IT environments for example caused by many legacy systems and/or an extensive reporting flow. A general rule of thumb is that the more End User Computing within a flow, the more complicated it is to capture and maintain.

Solid Data Management is a good starting point

To address data driven challenges in general, organizations adopt, develop and incorporate comprehensive data management ([Voor13]). The adaptation of data management for the complete value flow has led to the development of a set of measures supporting control of data quality from source to reporting. These measures support requirements of data driven regulatory compliance:

- *Data policy and organization*: this set of measures consists of a data management strategy, resulting in policies and guidelines. The set of policies and procedures that determine the who, how, and why of data management within the organization, thereby offering clear guidance on how data is governed and managed;
- *Data*: this set of measure consists of an end-to-end overview of data used (i.e. a data flow), specifying data sources, key data and meta data such as data ownership, characteristics, usage and modifications. As well as risks and limitations concerning the data;
- *Data processes*: this set of measure consists of data management processes (i.e. data lifecycle processes) including relevant external and internal data interfaces, requirements and controls. It includes data quality management and data issue resolution processes (all activities and procedures that aim to avoid errors or omissions in data and if errors are discovered, all activities to correct and prevent recurrence), data delivery agreements and/or service level agreements should be in place;
- *General IT*: this set of measures consists of a complete overview of the IT infrastructure landscape and the related – risk based – IT general controls, thereby safeguarding continuity and integrity. The landscape includes all relevant outsourced and/or managed by third parties systems and applications;
- *Application systems*: this set of measures consists of a complete overview of all used application systems: source systems, end user computing, risk engines and other tooling. This includes all descriptions concerning the functioning, controlling and continuity application controls, risk classification, access management, change and version management;
- *Data controlling*: this set of measures consists of all aspects that monitor and control the effectiveness of policies, procedures, processes, IT and application systems maintain the required data quality standards.

These measures to support capturing and maintaining the data flow within the reporting chain are heavily interlocked. For example, once a data quality rule has been defined, data ownership needs to be in place to validate that rule. Determining controls within a data flow, means that system owners need to be in place who can rely on risk analysis and policies. Having a data management organization (DMO) in place means pro-active and consistent governance of the data from record to report. An extensive elaboration on DMOs and data governance can be found in [Staa17]. So having data management in place

Organizations have difficulties in distinguishing a process flow from a reporting flow, assuming that existing process flows can function as a data flow overview

In order to fully understand and communicate the data flow to different stakeholders, it needs to be distinguished at different levels of granularity

supports capturing, sustainable maintenance and improving a data flow.

Principles for good future-proof Data Management

To define and manage data within the reporting flow, all the measures as mentioned before need to be placed. This makes data management complex to address. So organizations are especially interested in guiding principles to be able to cope with data (quality) challenges within the reporting flow from record-to-report.

- Organizations need to understand, identify and document how data moves (flows) and transforms throughout their reporting chain from source to reports. In order to fully understand and communicate the data flow to different stakeholders (e.g. report owners or an external supervisory), it needs to be distinguished at different levels of detail and granularity. The starting point is to identify systems, applications and databases in which relevant data has been stored (i.e. the system level). This overview subsequently enables the identification of relevant data sets and moreover how data sets move across the systems, departments and processes (i.e. data set level). Finally, each data set consists of data attributes; this is the lowest level of detail. Tracking and tracing data attributes from source to report is considered as data lineage. Data lineage is the most detailed description of the data flow, from the source system to its destination, including all transformations, mutations and enrichments it undergoes along the way.
- In theory, most organizations strive to completely capture the data flow in their IT systems by means of Straight-Through Processing (STP). In practice and for most organizations, data flows are (to an important extent) manually transported, transformed and controlled. Manual activities are usually time consuming and have an increased risk of errors. The current rise of software robotics offers a relatively low-cost alternative for automating the flows and reporting deviations, at least as a short-term solution until STP is embedded.
- Once relevant data attributes have been identified throughout the dataflow (data lineage), data can be classified into various categories. Classification of data enables categorization of data based on homogeneous characteristics in order to assess the impact and materiality of data elements in end-user reports. This also helps identify Key Data Elements (KDEs) which are the basis for reports.
- The next step is to document consistent definitions of those data elements. A data definition should explicitly describe the meaning of the data element and the context for which (business) purpose the data is being used. Data definitions should be documented in a centralized repository such as a Data Directory. A Data Directory can be considered as an inventory that specifies (e.g.) the source, location, definition, ownership, usage, and destination of all of the data elements that are stored in a database.
- When data definitions are formulated and documented, data quality criteria (also known as business rules) should be created from a business perspective. Data quality criteria can be distinguished in different dimensions (e.g. to measure the completeness, accuracy, timeliness, correctness or uniqueness of the data). Again, the set of applied data quality criteria should be stored in a single repository such as a Data Directory. Within the complete dataflow, risk based controls need to be in place. This consist of both application controls, manual controls as well as IT General Controls.
- The data quality criteria can subsequently be used to measure, monitor and demonstrate the actual quality of the underlying data elements which are used in your reports (finance, risk, management information). Dashboarding or data quality tools can support this process. Data which does not comply to the data quality criteria can be considered as data issues. A data cleansing process should be in place to cleanse those identified data issues.

Turning theory into practice: examples of practical approaches

Several organizations have defined their data flow based on the data management methodology and guiding principles as described above, some generating regulatory compliance in the process and setting up a management organization to maintain their data quality from source to report.

Insurer and Solvency II regulation

A large Dutch insurer pursued regulatory compliance for Solvency II. As Solvency II requires that each insurer has to 'prove that they are in control of the data which is used for regulatory reporting' the company recognized data quality and data management as a substantial domain of their SII approach. The current status of their data quality as well as the data flow/lineage was not clear. Nor was a governance body in place to maintain and assure sustainable data control.

The approach consisted of setting up a data governance, including a Target Operating Model for the Data Management Organization, standardized data life cycle & governance processes, policies, roles and responsibilities. Simultaneously the insurer assessed the data quality of SII relevant master data. providing data quality insights at an early phase of the implementation is beneficial for the common understanding of data, data quality, the requirement to address data at the smallest detail level (Key Data Elements) as well as generating a changing attitude towards data. Visualization of data quality makes people from executive level (e.g. the data owners) through to the operation level (data entry) understand how good data quality and consistent data definitions impacts their daily business as well as chain overarching processes. This attitude supported a speedy design of complete and a clear end-to-end description of their data flows within the reporting chains, including the extensive usage of End User Computing (here: MS Excel) within their actuarial processes. As an additional benefit of these insights into their data flows, discussions started to further improve and automate the processing of data, mainly in the actuarial departments.

Data quality online investment bank

This bank strived to become a data driven organization, where it had – as Tier 2 bank – less focus on regulation and more on fact based control information and fast customer insights. For these purposes they defined two tracks:

1. realizing a data management organization based on data governance, process and IT controls and retention framework (e.g. privacy retention periods);
2. setting up a data flow for reporting purposes based on new functionalities of a data lake (see also the 'Data Lake' text box).

Within this data lake quality criteria and definitions for data attributes where defined for both input from internal and external data suppliers as well as data users (i.e. data scientists). This meant that within the data lake attributes needed to be known and governed, based on the set of – interlocked – measures for data quality in the reporting flow.

So, whether the data in a reporting flow is governed from a compliance or innovation perspective, data quality measures do always need to be in place.

Data Lake

There are two primary solutions for storing large amounts of data for analysis purposes: a data warehouse and a data lake. While they serve the same purpose, they differ in some key areas ([Kuit16]).

A data warehouse is a combination of multiple databases and/or flat files, creating one integrated, time-variant and non-volatile collection of data. In practice, this means that data from multiple databases is stored in the warehouse through an ETL (Extraction, Transformation, Loading) process. This process ensures the data is integrated before entering the data warehouse. The fact that it is time-variant means historical data is stored in the data warehouse where the data does not change once it's inside the data warehouse.

Recently, the concept of the data lake has made its appearance. The idea of the data lake is similar to a data warehouse: providing a large collection of data to analyze. However, whereas the data warehouse only uses structured data, the data lake uses a combination of structured data and unstructured data (emails, social media, PDF files, and so on). As it uses this combination, the data is not integrated before entering the data lake. Rather, the structure of the provided data is determined when the analysis starts. This also means a data lake is more agile in its configuration than the data warehouse: it can be reconfigured as needed.

In practice, both solutions affect the insight an organization has into its record-to-report process differently. As a data warehouse has a more rigid structure, it should theoretically be quite easy to see where the data used for the report originates from. In practice this is not always the case, as organizations do not always have complete insight into their data warehouse structure and the ETL process, which were often built relatively long ago. The data lake offers better insight into the record-to-report process, as the analysis (and thus the structure) is very flexible and created only as early as it is needed for the report. This means that as you create the analysis for the report, you simultaneously define the data flow.

References

[Jonk12] R.A. Jonker, *Datakwaliteitsonderzoek*, Compact 2012/2.

[Kuit16] P. Kuiters, M.A. Baak, *Waarde genereren uit klantdata vraagt om meer dan slimme techniek*, Compact 2016/4.

[Staa17] A.J. van der Staaij, J.A.C. Tegelaar, *Data management activities: centralize, de-centralize or best-of-both-worlds?*, Compact 2017/1.

[Voor13] M.A. Voorhout, A.J. van der Staaij, S. Swartjes, *Masterdatamanagement: van frustratie tot totaalaanpak*, Compact 2013/3.

[Voor16] M.A. Voorhout, C. Cornelissen de Beer, R.A. Jonker, *Masterdatamanagement: de verandering centraal*, Compact 2016/1.

About the authors

P.A. Rothwell is a partner at KPMG Management Consulting in the Financial Management group. He has more than 15 years of experience in the financial services and operates at the intersection point between Finance, Risk and IT. Following a background in data analytics, financial IT systems and reporting processes, Paul has led various banks and insurance companies through a finance transformation process. Lately, Paul has been engaged in various programmes related to data-driven finance and Finance and Risk Data Management.

R.F. van der Ham is a senior consultant at KPMG Enterprise Data Management. He is primarily involved in the financial services sector, focusing on issues around regulatory compliance, data governance and data quality. He has been involved in numerous cases regarding compliance and data quality within the financial services sector as well as cases on MDM and control within the industrial sector.

M.A. Voorhout is a senior manager at KPMG Enterprise Data Management. She has been involved in numerous, international data management cases. She has a background in both Master Data Management and Enterprise Content Management and focuses primarily on data-related issues regarding regulatory compliance within various sectors.

D.A. Foudraine is a consultant at KPMG Enterprise Data Management. He focuses on issues in the areas of data quality, governance and has been involved in cases regarding regulatory compliance, governance, and data quality in the financial services sector. Along with this, he also has experience as a business integrator in the retail sector.