Deloitte.

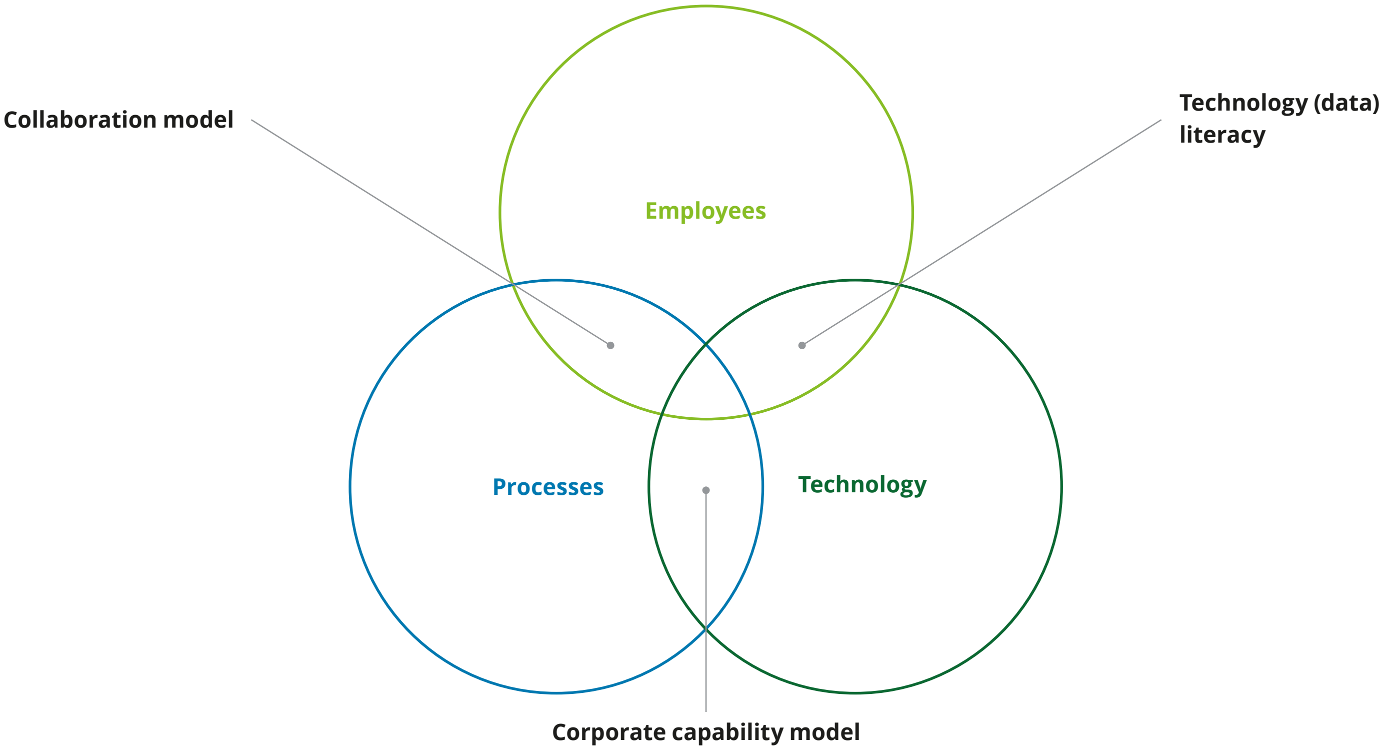
Data Quality and Data Meaningfulness in their Corporate Context

January 2023

Introduction

Modern enterprises leverage data and AI to keep up with the high dynamic of an ever-changing business world. On the way to achieve that goal almost all of our clients face the same challenge: To transform their strategy, operating model and way of doing business such that data is used to drive the corporate capabilities used in processes that generate value.

One of the recent concepts of achieving this is the data mesh, that shifts responsibilities towards the business to produce data products that are standardized in their creation as well as in their access. The data mesh paradigm, however, is not a purely technological invention, but rather a socio-technological endeavor that brings together technology, employees, and corporate processes. The classical Venn diagram of these three is given below. We have added the usually omitted descriptions of the links between the adjacent pairs: Employees possessing an intuition for the capabilities and limitations of technology ("technology literacy"), the enterprise being able to use technology in processes ("capability model") and the employees working together along the process landscape ("collaboration model").



Achieving literacy, data capabilities and collaboration are at the core of a successful transformation. Here, Deloitte presents a series of papers that explain key aspects how to achieve these three along the transformation towards becoming a data driven enterprise. The series is structured into strategic, tactical and operational aspects of data driven work.

Beginning with the strategy framework we are working along, we introduce our orchestrator for the data transformation journey. As the major tactical pillars of the transformation we focus on the required governance as well as the data-centric process landscape in two further articles.

These concepts are underpinned by operational tools such as data catalogs, data quality and IT platforms which we are also covering in an article. Since these developments need to be sustained by specialized change management, a separate article is dedicated to this topic.

The journey to a data-centric enterprise is a complex transformation that continues to bring new challenges and insights. We will continue to expand and add to our series of articles.

Data Quality and Data Meaningfulness in their Corporate Context

The transformation towards a data-driven business is frequently centered around a more fruitful use of data. One common concept of achieving this is the data mesh, storing data close to its source and shifting responsibilities towards the business to produce data products that are standardized in their creation as well as in their access. To ensure trust and reliability towards the published data products, data quality is a central requirement. Nonetheless, we have seen in many projects that for our clients the precise meaning and implication of 'data quality' remains unclear.

This paper aims at providing a holistic view of data quality considering strategic, tactical, and operational aspects and considering both IT-side and business-side responsibilities (see below figure). Readers with a strong business and management background will most likely gain most from the first page of this paper, while adept suppliers (e.g., data engineers) or users of data (e.g., analysts or data scientists) may gain most from the second and third pages.

The top-most strategic layer emphasizes the creation of business value by means of enhanced data quality, the corporate (data) strategy thereby aims at combining IT and business responsibilities regarding data towards corporate goals. As a reader working on data quality from a strategic point of view it is therefore imperative to define how data quality is part of the corporate strategy and structure the collaboration between business and IT.

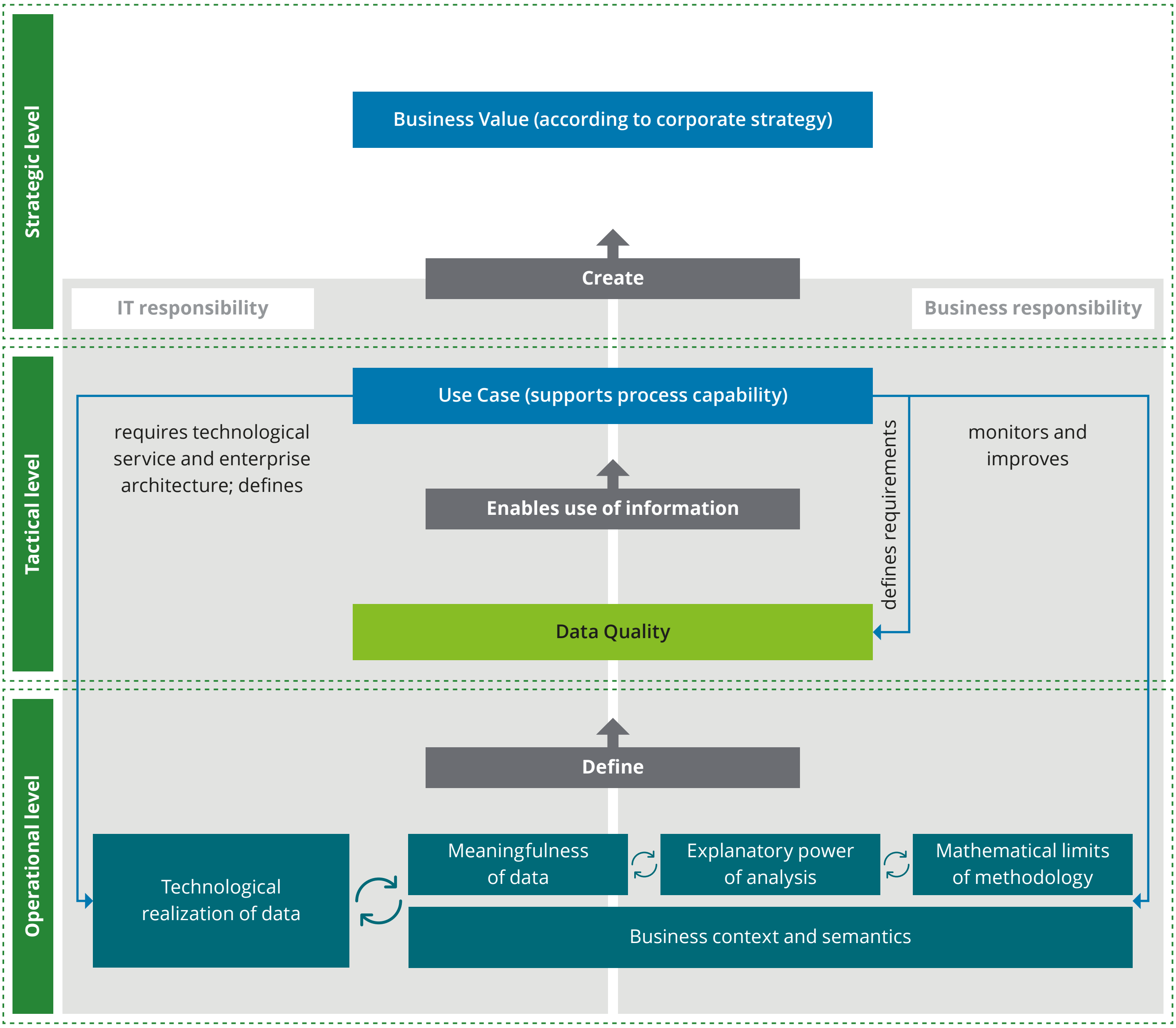
The tactical level is significant for bridging the gap between the strategic business value ambition and the operational handling of data. The key assumption here is that tactical work on data-driven use cases supports business capabilities that are used in business processes. Only utilizing use cases after their implementation creates the business value that was promised on the strategic level. Consequently, data quality does not only have a relevance as a strategic goal, but needs to be managed on the tactical level, so that use cases can consume information in the relevant context operationally. As a reader working in middle management your main responsibility lies either within facilitating the exchange between use case development and data quality requirements, enriching knowledge of data from your use case experience, or updating technological and architectural requirements to make use cases work.

To implement use cases there are certain implicit requirements towards data quality that may be updated with time: while the information required for a use case needs to be complete, the data itself does not need to be and, depending on the use case, can still contain missing values (e.g., identification of machine defects requires absence of data when the machine is broken; the absence of data is the information needed to identify that there is a defect and, hence, depending on the context missing values may not be a data quality issue).

On the operational level, working on a use case and understanding the relevant data has three distinct components. First, it requires a certain technological proficiency, secondly there is a specific need for IT services to provision results and finally both of this requires an enterprise data architecture leading to guidelines and principles to design a solution architecture. Therefore, the way data is stored depends on the architectural paradigm a company may employ. In addition, while some data is stored in tabular form, certain applications may work best with systems that store data in other form (JSON, XML, documents, graphs, and more). Hence, the type of utilization dictates the technological realization of data in systems. When it comes to the technical characteristics of the data platform, the responsibility for storage (e.g., setting data types), scalability, availability, and service deployment is on IT-side.

Consequently, the technological landscape must be designed in a way that root cause analysis of possible data quality issues is as easy as possible. Data lineage and its visual representation help identify risks in the data flows and traceback errors. Such representations are frequently annotated by metadata about the origin of and changes to data as well as further attributes illustrating its propagation process (data provenance). These metadata and further automated unit tests provide information about the correctness of data and the transformation logic making it simple to identify technological data quality issues.

Nonetheless, as mentioned above, data quality consists not only of technological aspects, but depends on context and business understanding. Data catalogs (providing metadata), master data models, or ontology tools (representing abstract relations between entities) are frequently used to make the relation between technology and business context more tangible. In addition, they complement integral capabilities of technological services such as data profiling, validation and or cleansing by conveying additional information about data transformation logic.



The necessary condition for data-driven use case work is taking the context in which data is used into account. The criteria can differ per use case and hence may influence the quality level of the same dataset in contradicting ways. Conversely, the characteristics mentioned above have been considered a sufficient condition in the beginning of the 2010s data science hype cycle. Data (e.g., numbers) obtain meaning in the context of the other available information (e.g., other columns in a table): The number 1.95 could be anything from a currency conversion rate to the break-down threshold of a machine and the column header and other columns clarify this issue. On the other hand, the number 1.95 can never be a denomination Japanese currency, Yen, since it does not possess decimals. Knowledge of how currencies work, their semantics, clear up this matter, but it is not part of the dataset.

Therefore, data is more or less useful depending on the context in which it is stored and correct or not depending on the semantics of their use. A common technique to capture such semantic knowledge is the use of ontologies. Ontologies are formalized descriptions that capture relations between business entities and their abstract realization as data. Adopting ontologies inherently builds knowledge graphs as descriptive tools which have proven useful in data product creation, particularly when connected to a data catalog to provide a holistic understanding of data from all perspectives.

As a result, semantics and context are two complementary aspects of data quality. Both contribute to what is called the 'meaningfulness' of data, be it the common-sense usefulness for an analysis (e.g., by means of an adequate level of aggregation) or the quality of statistically sophisticated 'feature engineering' used in data preparation for machine learning.

The meaningfulness of data is intimately linked to their use in an analysis of any kind. The type of analysis in turn determines the meaningfulness and explanatory power of the result. In addition, an analysis can provide insights by itself that does not necessarily need to be obvious from the data (e.g., meaningful customer clusters are derived using data mining techniques that use relevant data, but the clusters themselves are not evident in the raw data). This has two implications: On the one hand, the methodology of analysis must conform to the problem statement and the characteristics of the available data. On the other hand, more complex approaches to answer a business question may prove to be more insightful but are associated with stronger assumptions that could limit the analysis' general usability. Consequently, the type of analysis and its mathematical basis and assumptions limit this quality: The explanatory power of an analysis is determined by the meaningfulness of the data used, by the fit of the methodology to the problem and by the mathematical assumptions and limitations of the statistical tool.

Naturally, IT responsibilities are closely connected to business requirements. Therefore, the obligation to ensure meaningfulness of data is shared by IT and business. However, while the technological implementation of data in IT systems can frequently be ensured by automatic means of control (e.g., alerts that are fixed in a ticket-based system), the business' responsibility for the explanatory power of analyses and adherence to the restrictions imposed by the analysis methodology is mostly not that straight-forward. Particularly, lack of data quality or quality of an analysis' results may have multiple qualitatively different reasons:

1. The business processes or business model may have changed (e.g., by introducing new products or changing the sales structure)

2. The customer's behavior could be different (e.g., as a consequence of changes in customer behavior, or in more obscure cases such as fraud)

3. The environment in which the interaction of company and customers takes place could have been affected (e.g., by new regulations or legislation, a financial crisis or a pandemic)

4. The data are in fact correct, but the analyst's understanding of the data is not.

Considering this spectrum, the root cause of the perceived data quality issue may be unclear and requires considerable domain knowledge, conceptual insights and mathematical skill. The business-side responsibility for data quality is to ensure that these four seemingly conflicting types of data quality impediments are discerned and resolved. Particularly, tactical work on data-driven use cases based on data products that fuel the strategic business value improves and continuously monitors knowledge of the business context to sustain data quality from the business side. Then, data is not only explained in its technological and governance setting, but in its business context. Only then, data genuinely provides value.

Contributors

Mark Arndt
Consultant
Strategy, Analytics and M&A
+49 151 1917 5124
maarndt@deloitte.de

Dr. Max-Hendrik Böttcher
Manager
Cyber & Strategic Risk
+49 151 5448 3572
maboettcher@deloitte.de

Anica Buchholz
Manager
Human Capital
+49 151 5448 4563
abuchholz@deloitte.de

Thiemo Eberle
Senior Consultant
Enterprise Performance
+49 151 1829 5033
teberle@deloitte.de

Frank Eisenhauer
Director
Strategy, Analytics and M&A
+49 151 5448 3868
feisenhauer@deloitte.de

Dr. Christoph Euler
Senior Manager
Strategy, Analytics and M&A
+49 151 1268 3336
ceuler@deloitte.de

Markus Hafner
Senior Consultant
Strategy, Analytics and M&A
+49 151 5448 4069
mhafner@deloitte.de

Kai-Uwe Hess
Partner
Enterprise Performance
+49 151 1829 4406
kahess@deloitte.de

Margaretha Hirsch
Senior Consultant
Strategy, Analytics and M&A
+49 151 5448 4071
mhirsch@deloitte.de

Paul Holbein
Senior Consultant
Strategy, Analytics and M&A
+49 151 5807 6346
pholbein@deloitte.de

Martina Hornung
Partner
Human Capital
+49 151 5807 3900
mhornung@deloitte.de

Dr. Benjamin Klör
Manager
Enterprise Performance
+49 151 5448 4579
bkloer@deloitte.de

Dr. Sebastian Olbrich
Partner
Strategy, Analytics and M&A
+49 151 1488 0375
solbrich@deloitte.de

Ralf Pierson
Senior Consultant
Human Capital
+49 151 2103 3858
rpierson@deloitte.de

Christian Roese
Senior Specialist
Strategy, Analytics and M&A
+49 151 5807 6239
croese@deloitte.de

Marco Ruetten
Manager
Enterprise Performance
+49 151 5807 5861
maruetten@deloitte.de

Jan Scherpinski
Senior Manager
Strategy, Analytics and M&A
+49 151 5448 3411
jscherpinski@deloitte.de

Friedemann Schestag
Consultant
Strategy, Analytics and M&A
+49 151 1268 3791
fschestag@deloitte.de

Tim Staab
Consultant
Strategy, Analytics and M&A
+49 151 1488 1747
tistaab@deloitte.de

Julian Steiner
Senior Consultant
Human Capital
+49 151 5807 4575
julsteiner@deloitte.de

Aizhan Suleimenova
Consultant
Strategy, Analytics and M&A
+49 151 5448 3524
asuleimen@deloitte.de

Tobias Wätzig
Senior Manager
Strategy, Analytics and M&A
+49 151 5800 1565
twaetzig@deloitte.de

Marc Willemsen
Consultant
Strategy, Analytics and M&A
+49 151 1488 0730
mawillemsen@deloitte.de

Jan Wohde
Consultant
Human Capital
+49 151 5448 4643
jwohde@deloitte.de

Contacts

Frank Eisenhauer
Director
Strategy, Analytics and M&A
+49 151 5448 3868
feisenhauer@deloitte.de
Frank is a Director in Deloitte's AI & Data offering with a focus on Data-Strategy, Data-Transformation and Data-Management. He designs and implements data-driven-transformations with an objective on data-value-creation with a holistic and comprehensive approach. In his opinion, data transformation is a team sport.

Dr. Christoph Euler
Senior Manager
Strategy, Analytics and M&A
+49 151 1268 3336
ceuler@deloitte.de
Christoph is a data strategy consultant and data scientist, combining the two in his endeavor to help enterprises transform into data-ready organizations. He has a particular focus on the CEO and CDO organization's operating model design, data / machine learning governance, collaboration models and data literacy. In his opinion, data is a people business - technology is more readily available.

Christian Roese
Senior Specialist Lead
Strategy, Analytics and M&A
+49 151 5807 6239
croese@deloitte.de
Christian is a senior specialist lead in our Analytics & Cognitive Offering and has over 16 years of experiences of data management consultancy Christian is part of Deloitte Consulting since October 2018 and supported in his career mainly banking/ insurance sector clients for digital transformations. Since 2004 he was responsible for various data management initiatives for big data, data warehousing, data quality or data governance up to the level of program manager and solution architect.

Glossary

Data Mesh
The data mesh is a domain-driven socio-technological approach for creating decentralized data architectures. It is based on decentral governance structures as a foundation for generating sustainable business value using standardized and re-usable data products. It relies on a flexible collaboration model accross the entire enterprise.

Data Product
A data product is a set of data that is made available for the usage of employees or systems via a standardized API on a marketplace. Its purpose is to realize use cases and therefore to enable the implementation of data-driven services.

Data as a product
Synonymous to Data Product.

Use Case
A use case creates business value by fulfilling an explicit objective. Use cases are based on existing Data Products.

Data Catalog
A data catalog is the central inventory for all data assets within the company. It is made understandable via a glossary of frequently used terms and by highlighting the technical and business data lineage as well as transformation logic.

Data Governance
Data Governance is the discipline that connects data processes, and corresponding roles and responsibilities by formulating binding enterprise-wide policies.

Ontology
Ontologies are formalized descriptions that capture relations between business entities and their ab-stract realization as data.

Data Domain
A data domain takes ownership of data relevant to a common area of interest and implements roles that are responsible for expanding and maintaining the usability of this data.

Designed by CoRe Creative Services. RITM1180848