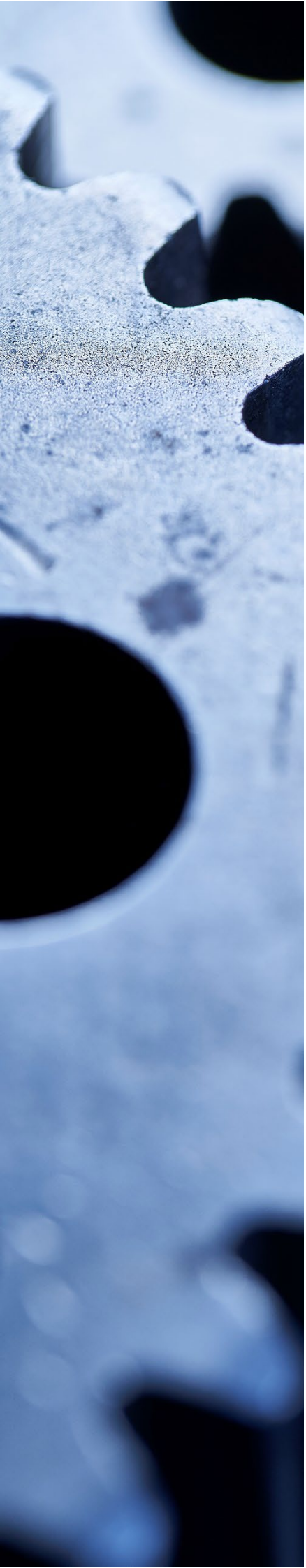
Data Quality by Design

How to Transform Data Policy into Data Operations

WHITE PAPER

datasynthesis



Contents

datasynthesis

Introduction

Data is the lifeblood of financial markets. Whilst the importance of data is freely acknowledged by the vast majority of financial institutions, historically this understanding has not translated easily into action and practice. Take the $400 million fine[1] recently imposed on Citi by US regulators for longstanding deficiencies in its risk management and data governance programs. Other fines to other firms will no doubt be forthcoming. Whilst many firms now have a Chief Data Officer (CDO), half regard their role as "nascent and evolving" and only around 30% have a well-articulated data strategy in place[2]. And all this is against a backdrop of exponential growth in data volumes, ever-more rigorous enforcement of an expanding number of regulations, and competitive pressure on firms to use data and data science as sources of new business and innovation.

Given these pressures, then just why has it proven so difficult for the majority of institutions to achieve enterprise-level control of data and data quality? History and culture have been a big part of the problem. At many financial institutions, existing data management tools were often commissioned as point solutions for specific departmental data issues. Given this context, it is unsurprising that these tools lack the capabilities necessary to support modern requirements for enterprise-level data governance and data quality. Achieving enterprise goals using such isolated and fragmented tools requires an enterprise-wide integration effort that is enormously complex, costly and resource-hungry. Such an approach has shown itself unable to keep pace with changing business needs and regulatory requirements for data.

"…WHEN NEW BUSINESS OR REGULATORY REQUIREMENTS ARE IMPOSED, THE BUSINESS CAN DEFINE CENTRALLY WHAT DATA CHANGES ARE NEEDED AND DEPLOY THEM INTO DEPARTMENTAL OPERATIONS WITHOUT CODING…"

A different approach is needed. Business staff responsible for data should be able to implement data policy without a mutually draining dependency on IT staff. A key design goal should be to use modern technology to enable the business to directly *transform data policy into data operations*. For example, when new business or regulatory requirements are imposed, the business can define centrally what data changes are needed and deploy them into departmental operations without coding and dependency upon departmental IT resourcing. The business becomes the owner of a dynamic but formalized data manufacturing process, one that delivers business agility, enforces top-down policy into operations and proactively embeds quality in all flows of data.

Such ambitious goals could not be achieved without acknowledging and respecting an institution's existing data environment. So let's start by exploring the detail of how people, business and technology have influenced enterprise data quality and why it has proven such a difficult goal to achieve.

[1] See https://www.occ.gov/static/enforcement-actions/ea2020-057.pdf
[2] See https://www.newvantage.com/thoughtleadership

Data and Business Evolution

Close to Data and Caring for it

Over the past 50 years, every decade has thrown up new business opportunities in financial markets – there was always a new product to trade, a new venue to trade it on, or a new customer to trade it with. With every opportunity, there were people around who were smart, hardworking and motivated, and who happened to have the additional advantage of being in the right place at the right time. Often a small team of one or two key staff developed the trading system for the business, often based on desktop databases and spreadsheets. Regardless of architecture, the key to why data quality was not typically an issue at this stage was *people*; the small number of people, all highly knowledgeable of the life cycle of data they used, and motivated to get any issue resolved. The people involved care about the data, they were close to it and were incentivised to ensure it is fit for purpose.

Technology as a Business Barrier

As the business succeeded, expansion followed. To enable the business to scale, more staff were hired and the original trading system and spreadsheets were upgraded to use more robust data tools, built by the new systems development team. Figure 1) below illustrates a typical data environment that evolved from these new initiatives, showing the complex data processing flows that develop as the business grows. Unfortunately, whilst more robust, these new data tools introduced a new operational barrier on the business. The trading staff who knew the data could no longer fix the problems they had identified. And the technology staff who could fix the data didn't know it as well, and any fix needed was a competing task amongst many other responsibilities. This mismatch of priorities, skills and responsibilities became a major source of frustration to the business and a real barrier to improved efficiency.

"THE TRADING STAFF WHO KNOW THE DATA CAN NO LONGER FIX IT…THIS BARRIER BECOMES A MAJOR SOURCE OF FRUSTRATION AND INEFFICIENCY…"

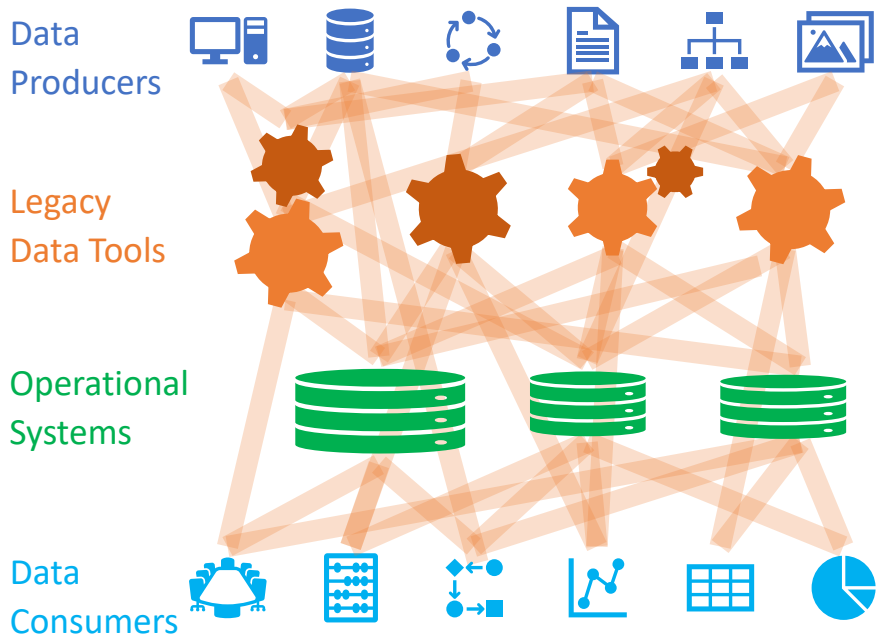


Figure 1) Legacy Data Tools in a Complex Network of Data Flows

Departmental Data is Enterprise Data

As growth in one department progressed, others grew in parallel along with centralised functions such as risk, compliance and finance. The relatively simple data environment of Figure 1) expanded and effectively became duplicated many times over. Cross-departmental co-operation and dependency, plus requests from central departments for information, started to introduce an enterprise context for data. This had painful consequences, as it was soon realised that departmental data only conformed to departmental standards, or indeed to no standard at all. Reconciliation between cross-departmental systems became a frustrating and unwanted drain on resources. And added to this was increasing external pressure from bodies such as auditors and regulators, who needed an aggregated, enterprise-level view of data.

Data and Technology Evolution

The business evolution of data quality described above provides some indirect but very significant reasons behind why the business use of data technology has not driven greater data quality. But the past evolution of data technology itself is part of the problem.

Fragmented Tools, Fragmented Process

Data management is often comprised of separate tools for building business glossaries, data dictionaries, data lineage, tools for integrating data, tools for creating "master" sets of reference data and tools for measuring data quality. Multiple installations of such tools are frequently found across multiple departments, often from different vendors and often specific to certain types or sources of data. Typically, each tool was brought in as a point solution to a specific business issue. Given this context, it is perhaps unsurprising that these tools lack the scope, scale and administrative capabilities necessary to support modern requirements of enterprise-level governance and quality. They were not designed with this context in mind.

"…AS A RESULT BAD DATA MAY HAVE ALREADY SPREAD DOWNSTREAM, REQUIRING TIME-CONSUMING FIXES AND INDIVIDUAL DATABASE REWINDS TO BE MADE ACROSS MULTIPLE SYSTEMS…"

Isolated Data, Isolated Rules About Data

For example, each operational system within the enterprise is likely to have associated ETL (extract-transform-load) tools, containing hard-coded rules that ensure that data is both valid and transformed into the format needed. This is fine if nothing goes wrong, however if errors are found then the batch-based nature of these tools means that notification only occurs once the process completes. Due to time pressures on process scheduling, then as a result this bad data may have already spread to operational systems further downstream. To remedy the problem, time-consuming fixes and individual database rewinds have to be made across multiple operational systems, potentially by multiple IT teams. As illustrated by the isolated data tools shown in Figure 1), enterprise-level changes, such as the adoption of new standards to improve data consistency, cannot be applied in one place. Rather, changes have to be implemented many times over, once for each individual data "cog" in each department. Such an uncoordinated approach is resource hungry, complex and becoming prohibitively expensive to implement and maintain.

Data and Regulatory Evolution

Whilst business growth and technology evolution are both key factors in attempting to achieve enterprise data quality, regulation has been the biggest external driver of enterprise governance and quality initiatives.

History

Given the amount of regulation in recent years, Table 1) lists a necessarily incomplete history of some key regulations and their effect on data governance and quality within financial markets and the wider economy. The table misses many major regulations (e.g. IFRS or FRTB) but is illustrative of how regulation is driving the need for enterprise governance and quality. Regulation typically drives improvement in data management indirectly, due to the demanding nature of the reporting requirements imposed. Some regulation goes further however, through direct and explicit prescription of data management principles and/or use of data standards. Whether implicit or explicit, the consequence in either case is that improvements in data governance and data quality need to be made.

Regulation	*Data Category*	*Year*	*Applicability*	*Description*
SEC 15c6-1	Reference	1993	US Institutions	Settlement to T+3 driving process automation
CAD	Historic Market	1993/5	Global Banks	VaR adoption requiring clean historic data
SOX	Spreadsheets	2002	US Businesses	Governance of end-user-computing tools
Basel II	Risk/Credit	2004	Global Banks	Addition of credit risk to Basel I requirements
Dodd Frank	Risk/LEI	2010	US Institutions	Risk reporting and LEI entity standard introduction
Form PF	Risk	2012	US HFs	Reporting requirements for asset risk exposure
BCBS239	Risk	2013	GSIBs	Principles for risk data aggregation
Mifid II	Trade	2014	EU Institutions	Trade/transaction reporting requirements
Solvency II	Risk	2015	EU Insurers	Governance of capital calcs and risk reporting
GDPR	Personal	2016	EU/Global	Privacy and security of personal data of EU Citizens

Table 1) Major Regulatory Influences on Data Management

Competitive Advantage

Regulatory compliance and data governance are often perceived as defensive initiatives. However, as organisations increasingly realise value from data, it is becoming more widely recognised that this perception has, historically, resulted in missed business opportunities. A financial institution that has trust in its own data capabilities will be able to comply with new regulation quicker and at lower cost than firms still struggling with legacy architecture. And these same capabilities also promote faster, better decision-making and foster business innovation around data. Regulation can be a burden, but if approached strategically it can be a catalyst for improved profitability and competitive advantage.

Design Principles

We have seen some of the influences on data quality and governance: people, business growth, technology legacy, new regulation and growth in data volumes and usage. So how best can an enterprise put all these elements together to build towards operational processes that ensure enterprise data quality? Making the vital assumption the business has already been convinced of the benefits of improved data quality, the section below sets out some of our design principles for *transforming* the process of data quality management.

I. Business Should Both Define and Execute Data Policy

In order to put business in control of the data it uses and produces, the business should define policy, standards and rules. Technology should be used to enable business to directly execute and maintain policy within data operations. In order to achieve data quality, IT departments cannot both control data and simultaneously be granted this control as a secondary by-product of their main activity. The historical assignment of responsibility for data, policy and quality to IT serves neither the business nor IT well. A shift in both responsibility and control of data is key to addressing this.

"…THE BUSINESS SHOULD DEFINE POLICY, STANDARDS AND RULES; TECHNOLOGY SHOULD BE USED TO ENABLE BUSINESS TO DIRECTLY EXECUTE AND MAINTAIN POLICY…"

II. Enable Data Policy to Directly Define Data Operations

In order for business to both define and reliably execute policy, policy should directly define data operations. Effectively, a much closer connection needs to be made between the business changing policy (e.g. defining quality rules) and the new policy being applied to operational systems. Policy definition should effectively become a "no-code" development language through which the business controls data. And so instead of governance, operations and quality being disconnected, they are synchronised: from policy input; through to day-to-day operations; through to the delivery of high quality data to end users.

III. Require Enterprise Scalability in Capacity and in Capability

Achieving enterprise data quality requires a technology architecture that can scale enterprise-wide in both technical capacity and functional capability. Scaling in capacity means that technology performance is no longer a constraint on the business. Scaling in functionality means that such an architecture is easily extensible and highly flexible, for example in that it can handle any type of instrument, across any business line. And if designed for full enterprise-level usage, such an architecture will drastically reduce the costs and resourcing needed to both achieve and maintain enterprise data quality.

IV. Respect the Existing Operational Data Environment

Most firms have a large legacy of data and systems, so successful execution of policy should minimise disruption to existing operations and be non-invasive; it should have zero infrastructure footprint; it should be incrementally deployable to minimise risk and demonstrate incremental business benefit. A solution is needed that leaves operational data stores where they are now; recent efforts have demonstrated that re-architecting all of your trading and operational systems around a central data lake has not been a panacea for successful data management.

Platform Design

How best could we build a business platform for achieving enterprise data quality, one that would help financial institutions transform the way they manage data? Let's start with the design principles we have just outlined and see how they can be delivered.

Formal Data Control, in Everyone's Language

A first step in achieving Principle I in enabling the business to both define and execute data policy is to bring all policies containing rules, access permissions and workflows under formal version control, building towards "data life cycle management" which can be thought of as the data development equivalent of the Software Development Life Cycle (SDLC). Once this core data control framework is in place, the second step is to remove the technological barrier that has prevented data experts from managing the full lifecycle of data, from data definition to deletion and all the steps in between.

"…MANY MORE NON-TECHNICAL USERS CAN SAFELY AND SECURELY CONTRIBUTE TO DATA QUALITY MANAGEMENT ON A DAILY BASIS, IN MUCH THE SAME WAY AS THEY WOULD ROUTINELY ANSWER AN EMAIL."

As a result, the graphical user interface (GUI) for such a platform must support the complete data life cycle, without the need for SQL or any other specialized programming skills to implement transformation and validation rules. Visual design and display of policy rules plays a key part in this, combined with the use of Natural Language Processing (NLP) as a mechanism to enable users to describe workflows, rules and using language they are already comfortable with, integrated into their day-to-day working environments. As illustrated in Figure 2) below, data agents can be used to expose core NLP functionality as chatbots into modern collaboration tools such as Slack or MS Teams. Using these tools, many more non-technical users can safely and securely contribute to data quality management on a daily basis, in much the same way as they would routinely answer an email.

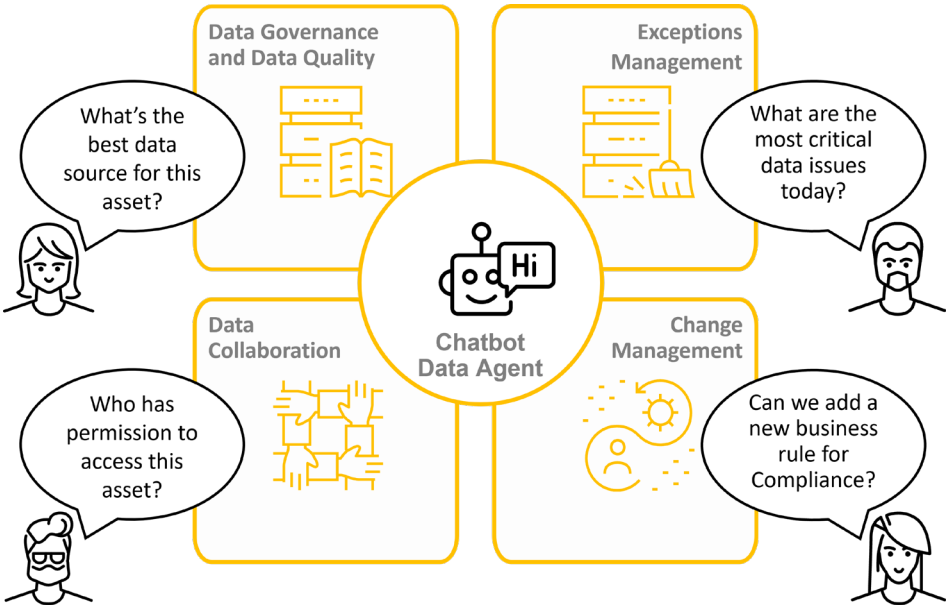


Figure 2) Everyday Language to Control Your Data Management Processes

Data About Data, Driving its Flow

Whilst Principle I deals with the importance of removing the technological barriers to data experts taking control of data, Principle II deals with the underlying issue of how data policy is implemented. So if data experts can easily define policy in terms of access rights, rules and standards, how best could such policy definitions be translated into what happens to data in day-to-day operations, and ultimately lead to improved enterprise data quality? Looking back at Figure 1), an alternative way of effectively asking the same question is how could enterprise data governance and enterprise data quality be directly connected into the cogs of departmental data operations?

Central management and co-ordination of legacy data tools is extremely difficult to achieve; so a different approach must be taken. The approach of aggregating a central repository of business rules is an established form of *metadata management*, used by many *data catalog* tools to allow users to understand the flow and usage of enterprise data. Metadata management is effectively a disconnected snapshot view of data flows in operations. Such approaches help in understanding how and where data flows, but they are not able to change the flow of data even if they have identified an issue.

Figure 3) Governance Policy, Directly Driving Data Operations, Enforcing Data Quality

"THIS METADATA ENGINEERING APPROACH TO DATA MANAGEMENT PRODUCES AN INTEGRATED AND INTERCONNECTED DATA MANUFACTURING PROCESS FOR DELIVERING ENTERPRISE DATA QUALITY."

What's needed is a more pro-active approach, where this central repository of business rules is combined with a distributed rule execution engine. This engine would take rules from the central repository, and transform them into machine-readable instructions, distributed and processed in real-time to directly validate and transform data in day-to-day operations. This *metadata engineering* approach to data management produces an *integrated and interconnected data manufacturing process* for delivering enterprise data quality, as shown in Figure 3) above. If enterprise data governance policy directly drives data validation and transformation in day-to-day operations, then enforcement of enterprise data quality standards logically follows.

Internet-Scale Modern Data Architecture

Principal III says that achieving enterprise-level data quality requires a technology architecture that can scale enterprise-wide in both technical capacity and functional capability. Dealing with this from a technical perspective first, then well-designed cloud-native applications, based open source software components within modern microservices architectures, offer effectively limitless and automatic scale in storage and real-time processing capabilities. Freed of legacy performance constraints, such an architecture can better reflect the full business process of data quality management, all within a single, multi-workload design.

"…A MODERN DATA ARCHITECTURE CAN AUTOMATICALLY SCALE TO MEET SERVICE LEVEL AGREEMENTS DEFINED BY BUSINESS NEED, RATHER THAN BUSINESS SCALING BACK ITS ACTIVITY TO AVOID BREACHING LIMITS SET BY TECHNOLOGY."

Combined with modern data type representation and data relationship navigation, the architecture can deal with all types of data and integrate many of the data management functions that traditionally have had to be implemented within separate tools. In summary, such a modern data architecture can automatically scale to meet service level agreements defined by business need, rather than business scaling back its activity to avoid breaching limits set by technology.

Incremental Implementation

Principle IV is an acknowledgement that the vast majority of institutions are not greenfield sites, and big bang deployment approaches to replacing legacy tools are at best risky, and in most cases completely impractical. As such any successful solution must be non-invasive in design. The proposed platform architecture outlined in preceding sections has many attributes that address this need. It is cloud native, enabling incremental deployment as a service with a zero infrastructure footprint and zero up-front capital outlay. The architecture is based on information about data (the rules) rather than the underlying data itself, and hence operational systems and data stores can be left in place. As illustrated in Figure 4) below, the rules themselves can be migrated incrementally. So starting with one system or department, rules can progressively be integrated within enterprise data governance policy.

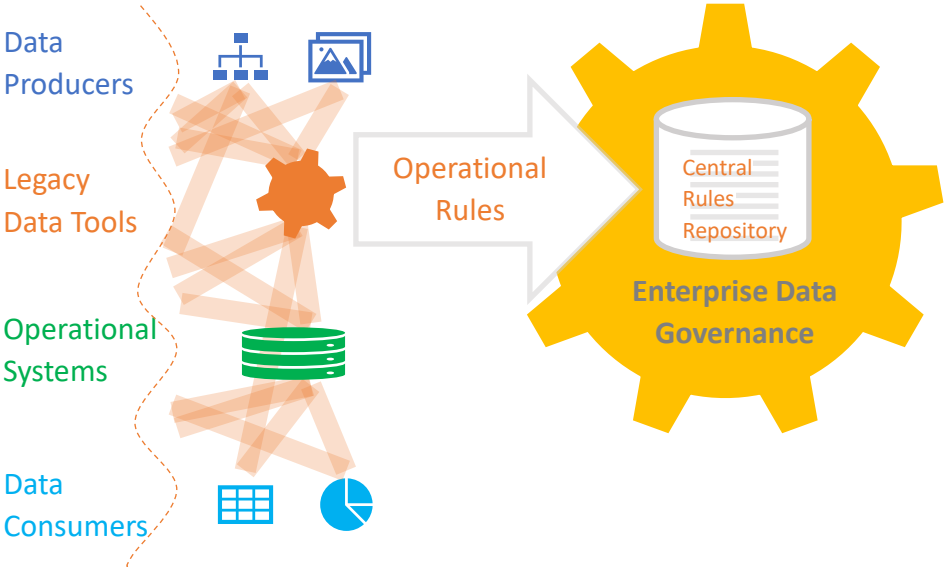


Figure 4) Incremental Migration of Rules and Integration within Data Policy

The Datasynthesis Platform

The Datasynthesis Platform leverages the experiences and principles described in preceding sections of this document to enable financial institutions to directly *transform data policy into data operations*.

Architectural Overview

The Datasynthesis Platform is unique in that for the first time it gives the business the capability to express data policies, standards and rules in plain English, and see data policy directly implemented in the flows of data around the enterprise. Given these data flows are driven directly by policy, any data quality breaks with policy are detected and staff alerted in real-time, preventing bad data from spreading to downstream systems. And to cope with these real-time and enterprise-wide performance demands, the Platform is cloud native in design and so automatically scales to meet any business workload. As shown in Figure 5) below, the Platform's data distribution capabilities simplify the flow of data between producers and consumers, leaving operational systems in place and greatly improving the efficiency of the entire data management process.

"…IT GIVES THE BUSINESS THE CAPABILITY TO EXPRESS DATA POLICIES, STANDARDS AND RULES IN PLAIN ENGLISH, AND SEE DATA POLICY DIRECTLY REFLECTED IN WHAT HAPPENS IN DATA OPERATIONS."

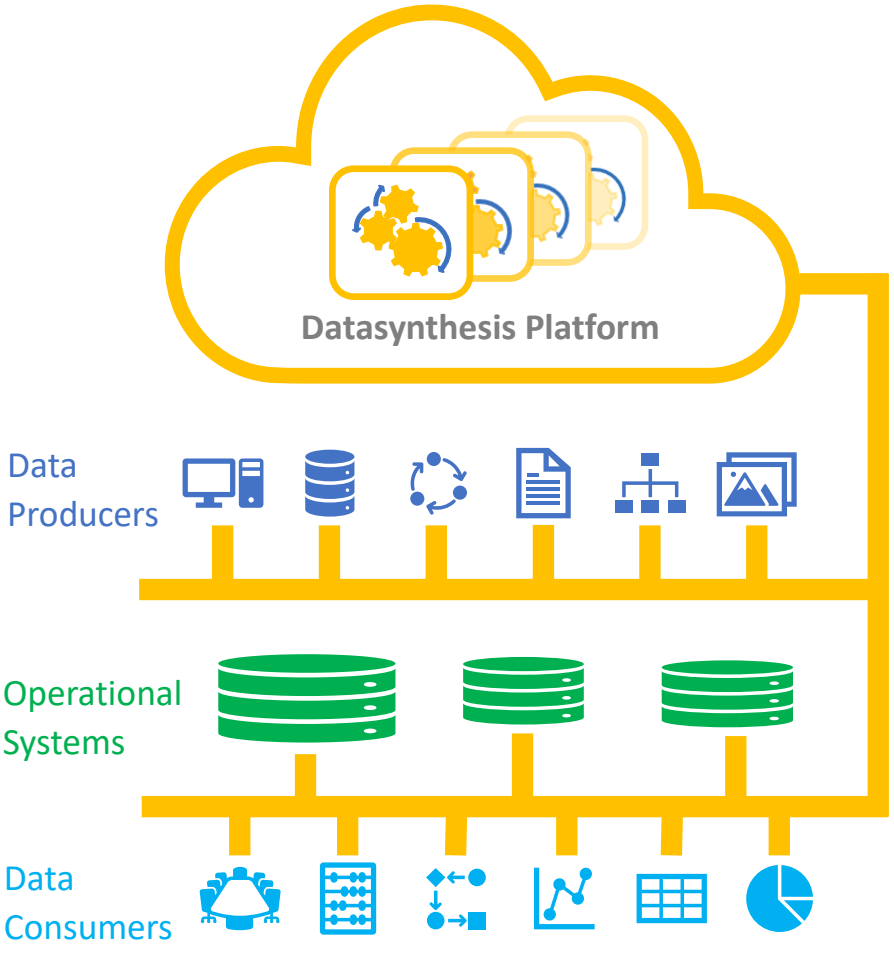
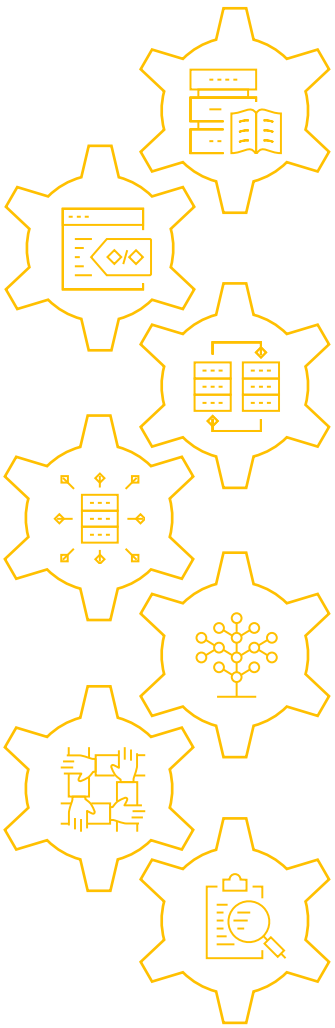


Figure 5) Modern Data Environment Enabled by the Datasynthesis Platform

Functional Overview

The Datasynthesis Platform delivers enterprise data quality through the loosely-coupled integration of the following sub-components of the data manufacturing process:



Data Governance – authorized data domains and source catalogs, data ownership, critical data elements life-cycle management, impact analysis when changes occur.

Metadata Management – glossary of business terms/meanings, taxonomies and rules, data-quality expectations SLA's, executable data policy operational metadata.

Data Integration – no-code modelling of conceptual and logical data meaning and cross-departmental relationships, supported by real-time data integration delivery.

Data Mastering – enabling multiple sources of data to be staged/promoted through different governance levels appropriate to user quality and governance needs.

Data Stewardship – data dictionary, data lineage, transformation rules, business rules to facilitate understanding of the data and related business expectations.

Data Collaboration – data workflows and exception management exposed into user productivity tools like Slack or MS Teams and augmented by Chatbot functionality.

Data Quality Management – near real-time data quality dashboards and KPI reporting across multiple critical data elements and systems of record.

The Platform brings the above functionality components together to form the data management equivalent of the Software Development Life Cycle, allowing data experts to plan, design, build, document, test, maintain and deploy the flow of data around the enterprise. So when new regulatory data requirements are imposed, the business can define centrally what data changes are needed and deploy them into departmental operations without coding or high dependency upon departmental IT resourcing. Those responsible for the data finally have the platform they need to transform their data policies directly into what happens in day-to-day data operations. And the whole business has an incremental pathway away from legacy data tools and environments to a simpler, low cost and highly agile modern data architecture.

Summary

We have seen how data quality issues have evolved at financial institutions. Business staff who know data best and use it day-in/day-out should be the people responsible for data quality. Unfortunately, the need to scale business through technology has introduced a barrier between data experts and technology experts who can change data, but are not responsible for it. This technology barrier has been exacerbated by the fragmented nature of legacy data management technology, and the way in which data management tools are isolated and managed only at a departmental level and scale. Achieving enterprise data quality based on such isolated and disparate tools requires an enterprise-wide integration effort that is enormously complex, costly and will never keep pace with changing business needs and regulatory requirements for data.

A different approach is needed. Using modern open source data technologies within a cloud-native microservice architecture, has the potential to free developers from the performance limitations of legacy data management tools. Whilst this is a huge and necessary step forward in technology design and business process implementation, if the process of data quality management is still highly dependent on IT, then nothing has changed for the business. We believe that achieving enterprise data quality is about *fixing the process of data quality management*, not the legacy approach of simply fixing data issues wherever and whenever they are found.

A fundamental step in designing for enterprise data quality is to enable *the business to both define and execute data policy*. Those business staff responsible for data policy should be able to implement it without a mutually draining dependency on IT staff. A key implementation constraint is that such a design be *non-invasive*. Implementation of such a platform should leave operational systems and data stores securely in place, whilst offering an incremental migration path to a modern data architecture. And building on modern data technology to enable scale in both capacity and capability, the final design aim is to *transform data policy into data operations*. Going beyond metadata management, we believe that *metadata engineering* is a key technique in achieving all of these aims.

In order to turn the above design principles into practical reality for our clients, we have built the Datasynthesis Platform. The Platform leverages modern data technology, no-code data collaboration and metadata engineering to enable:

- ✓ **The CDO to transform data policy directly into data operations**
- ✓ **Data experts to use plain English to control the data manufacturing process**
- ✓ **The CIO to incrementally migrate to a modern data architecture**
- ✓ **IT experts to spend more time on strategic delivery and less on fixing data**
- ✓ **Real-time prevention of data issues at source not manual fixes at many destinations**

If you would be interested in finding out more about Datasynthesis, the Datasynthesis Platform and how our approach to enterprise data quality could help your organisation, then please visit datasynthesis.com or email info@datasynthesis.com. Thank you for taking the time to read this paper!

About Datasynthesis

DATASYNTHESIS was founded in 2018 and brings to bear the multi-decades experience of the founders in designing, developing and successfully deploying commercial data management solutions for financial markets institutions. The company's mission is to enable data experts to take full control of enterprise data, though a no-code approach that enables greater participation by the business and greater collaboration with IT. The Datasynthesis Platform is an integrated data quality management platform based on modern, elastic-scale cloud technology, that delivers dramatic improvements in business efficiency and agility. See www.datasynthesis.com for more information.

Version 1.41