

Building Foundational Data Products for Financial Services: A MDM-Based Approach to Customer, and Product Data Integration

Ramesh Inala ^{1,*} 

¹ Data Engineer, USA

*Correspondence: Ramesh Inala (rameshhinala@gmail.com)

Abstract: Imagine a consumer financial services company with 20 million customers. Its sales and marketing organizations collaborate across product lines, deploying hundreds of marketing campaigns each quarter that aim to increase customer product usage and/or cross-buying of products. Each campaign is based on forecasts of customer responses derived from predictive models updated every quarter. The goals of these models are to achieve large return on investment ratios and to maximize contribution to local profit centers. What's important is that their modeling is based only on data created, curated and maintained by these marketing organizations. The difference today is that the modeling is no longer based solely on a small number of response-determined variables that are constantly assessed in terms of importance. A quarterly campaign update generates hundreds of statistical models — involving campaign responses, purchase-lag time, the relative magnitude of the direct effect, and the cross-buying effects — using thousands of variables, including customer demographics, life stage, product transactions, household composition, and customer service history. It's a network of models, not just a table of variable-by-residual importance values. But that's only part of the story of data products. The predictive modeling utilized by these campaign plans is based on analytics and data preparation, which are data products in their most diminutive form. These products would be even more elementary were they not crafted quarterly by highly skilled, experienced modelers using advanced software and processes. Most companies have enough data to create models that contain not simply hundreds of variables, but thousands, so that the focus can return to information instead of data reduction. These models largely replace the internal econometric models previously used to produce advanced forecasts in the absence of campaign modeling. People used these forecasts to simulate ROI and contribution forecasts for the planned campaigns. In the old days, reliance on econometrically forecast ROI-guideline contribution values reduced the reliance on the marketing campaign modelers because of a lack of trust in their predictive ability.

Keywords: Master Data Management (MDM), Financial Data Integration, Customer Data Platforms (CDP), Product Data Management, Foundational Data Products, Enterprise Data Architecture, Customer 360 View, Data Productization, Data Governance in Finance, Metadata Management, Data Lineage and Quality, Financial Services IT, Data-Driven Product Design, Unified Customer Profiles, Digital Core for Financial Institutions

How to cite this paper:

Inala, R. (2020). Building Foundational Data Products for Financial Services: A MDM-Based Approach to Customer, and Product Data Integration. *Universal Journal of Finance and Economics*, 1(1), 1–18.
DOI: [10.31586/ujfe.2020.1342](https://doi.org/10.31586/ujfe.2020.1342)

Received: September 30, 2020

Revised: November 26, 2020

Accepted: December 22, 2020

Published: December 27, 2020



Copyright: © 2020 by the author. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid advances in technology and data availability, matched with low innovation in delivery, have shifted the customer experience to the right, seeing demand for financial services evolve. Sensing these market forces, financial business leaders attempt to catch up in delivering service experiences that their consumers cannot do without — experiences that confuse and overwhelm with choice and simplicity at

inception but retain sticky habituation loops at their core. These trends have fostered an understanding on the part of business leaders that data and technology are no longer enablers for their strategic vision; data and technology have become their strategic vision. Banks and other financial service firms have to mottled customer experience pipelines that connect with all the consumer's motives and needs: motivation, inspiration, navigation, transaction, advice, and service to make data their sole focus of product offerings. Data is at the center of many business decisions in risk, treasury, customer, and operations; businesses should turn this knowledge to bring data closer to their customers [1].

Foundational data products are not end-user products but data enablers for use in transactions or as part of delivering agency, regulatory, and planning data services. AI may become a tool for automating basic processes that make successful customer experiences or allow different orchestration of existing product offerings that matter. Foundational data products think differently; they treat services and not technologies as the source of differentiation. A consumer may use transparency or algorithm knowledge to force their bank to provide similar pricing information, but getting insight on when to initiate a loan, anticipation of probable overdraft, help with initiating a loan application, or auto-executing enablement solutions integrated with data services: intelligent loan origination, savings product matchmaking, intimacy-based deal sourcing are the promise of data products [2].

1.1. Background and Significance

Over the past few decades, cost-effective data storage technologies and the rise of cloud computing made the storage and computational processing of vast amounts of data commercially feasible. In parallel, ever-evolving computational capabilities specific to the domains of artificial intelligence and data management enabled the democratization of Artificial Intelligence. With relatively small amounts of "good" and "enough" data, both new startups and technology giants have been able to successfully build products that leverage machine learning-enabled capabilities, such as image and speech recognition, translation and expert systems, recommender systems and game playing. These data-driven products become ubiquitous in a variety of contexts, providing significant value for everyday human activities. From exploring and making sense of information to engaging with customers to learning about new products and services, these products enhance the day-to-day digital experiences of consumers across diverse sectors, such as entertainment, gaming and travel, among others. As a result, consumers have come to expect similar levels of experiences in their digital interactions with their financial services providers, in areas such as customer support, product discovery, credit rating, analysis and mitigation of risks, and fraud detection. Until now, however, when compared to these other sectors, the design and the types of applications of artificial intelligence in financial services are relatively sparse. Yet, due to significant advances in the availability and the quality of financial transaction data, additional data-driven innovations can, indeed, take place in the sector [3].

2. Understanding Master Data Management (MDM)

The second section of this essay provides fundamentally important information that broadly defines and describes the methods and technologies of Master Data Management (MDM), and also offers insights related to the MDM use cases in financial services. MDM has been widely practiced in many business verticals, such as retail, energy and utilities, healthcare, etc. Understanding MDM techs and methods and their implementations in these domains helps inferring and refining its financial services implementations. Those financial firms with existing investments in cross-domain MDM techs and methods would achieve data product vision faster if they are encouraged to use proven MDM sector technologies and methods in supportive financial services implementation. Timely infusion of cross-domain MDM alternatives could even motivate the firms without any

cross-sector MDM investment to pursue the data product strategy. Because a financial MDM implementation that deeply leverages cross-domain pillars would have a better chance of succeeding.

Specifically, this essay will help ensure that an organization embarking upon the MDM journey grasp the foundational pillars and building blocks. MDM is also deeply impacted by organization-specific strategies, organizational design, policies and processes, etc. Hence, a one-to-one realization of such a tool-set is neither warranted nor might it work. Organizations should carefully consider MDM refined for their needs and strategies. MDM efforts are carried out as individual projects and as ongoing and support processes. MDM has multiple objectives and reasons to exist, some are exhaustive in scope and highly driven by regulatory requirements, while others are narrowly focused. The specific implementation and realization of the aforementioned pillars, building blocks and MDM characteristics at the technology and solution levels vary as well. The MDM landscape is dotted with a plethora of MDM solutions, whose end-state characteristics are not the same [4].

2.1. Research Design

To write this essay, we carried out a qualitative research using semi-structured interviews with MDM delivery executives of three organizations that have worked in MDM initiatives across various financial institutions. The objectives of the interviews were three-fold: (a) to explore the MDM constructs and metrics we had drawn from the existing literature, (b) to gather practical experiences about managing successful MDM initiatives, and (c) to examine the challenges related to MDM, and the solutions to overcome the challenges.

The interviews were conducted over phone and lasted about an hour each. Each organization provided us a unique view of MDM activities along with detailed implementations in different financial institutions. To minimize subjective bias, most of our interview questions were multiple-choice, with respondents asked to choose the answers that best represented their experience. The interviews were taped with respondent permission and transcribed later. Based on the research and feedback from the interviewees, we formulated the MDM enabling constructs and metrics that we had drawn from the existing literature, and summarized them in a structured textual format. We then sent them the structured format and asked them to rate those attributes on a 5-point Likert scale ranging from 1 (not important at all) to 5 (very important). Based on their feedback, we expanded or shrank the attributes as needed.

We also formed six enabling constructs for IT to manage and use, or consider using, as part of the product development and operations strategy for MDM components, and workflows for the enabling IT capabilities, as well as guidance for managing MDM-enabled business processes. This information should help decision-makers place more emphasis on the key constructs that are the main relevant enablers of MDM and the problem that is expected to be solved by organizations implementing MDM. Of course, this does not imply that they should completely ignore the other dimensions of MDM that are possible enablers.

Equation 1: Data Integration Efficiency (DIE).

$$DIE = \frac{E_{matched}}{E_{total} - E_{duplicate} - E_{conflict}}$$

Where:

- DIE : Data Integration Efficiency
- $E_{matched}$: Number of accurately matched customer/product records
- E_{total} : Total ingested records from all sources

- $E_{duplicate}$: Duplicate records
- $E_{conflict}$: Conflicting or unresolved records

3. The Role of Customer Data in Financial Services

To build foundational data products that will enable an organization to meet its strategic and tactical goals, we must begin with data that supports customer engagement and interfacing. Such data, referred to here as customer data, contains all the attributes and features resulting from an organization's activities to attract, engage, sell to, service, and keep a customer. These data include an account holder's account information, interaction, commerce, and complaint activity. They also include the risk, compliance, economic, and marketing segmentation attributes, as well as the credit scores of the consumer or small business. Given its importance, customer data for financial services has received a lot of attention. Yet, as with many things, attention has not necessarily resulted in success [5].

The increasing reliance on customer data for financial services to conduct their daily operations has not been equally matched by successful efforts to create, maintain, and pull insights from these data in a unified manner. Customer data is often developed, maintained, and made accessible as a by-product of these group-level activities, not as an integral part of the organization's business strategy or operating model. Equally important is that customer-level information is only as good as the quality of the data elements that comprise it, and such quality is only as good as the organization's commitment to preserve their integrity through organization-wide policies and procedures that govern interaction management, inquiry handling, transaction processing, record creation, and data storage. If business users sidestep formalized procedures, they create garbage-in, garbage-out customer data products that erode the trustworthiness and reliability of customer data across the organization [6].

3.1. Importance of Customer Data

In recent years, financial services firms, and retail banks in particular, have increased their investments in core customer data management systems, including customer reference data, customer due diligence and Know Your Customer data, and customer relationship management systems. Such investments align with the rapid consumer adoption of digital services, increased expectations for a seamless end-to-end experience, and the growth of big tech-enabled ecosystems that embrace the technology capabilities of digital native businesses. Together these shifts have created a focus on customers, with customer data solutions playing a key role in achieving fintech-like customer experiences.



Figure 1. Importance of Customer Data.

Improving customer experiences has become a top priority for many banking executives, with consumer satisfaction surveyed as the primary driver behind core banking technology investments. Banks aim to simplify customers' interactions from onboarding and servicing to understanding third-party data sharing and rewards – one experience, one line of sight into the customer's financial world at any point in time, one notification of required action, and one security checkpoint. Banks recognize that a positive and longstanding customer experience is key to attracting business customers, deepening those relationships, and unseating fintech that rely on offering only niche products as they compete to eat into the traditional business bank's market share through improved digital experiences. At the same time, rigorous KYC checks mandated by anti-money laundering and Know Your Customer laws are a requirement for compliance, protecting banks from the costs of fraud and dissatisfied customers from poor service. Improved core banking systems are necessary to reduce the cost and time of onboarding and servicing customers.

3.2. Challenges in Customer Data Management

A multitude of factors can negatively affect the accuracy of customer data. High levels of employee turnover, which create a fluid sales associate workforce, can hinder the success of customer onboarding and can result in inaccuracies in CRM data updates. Additionally, sales associates may not prioritize accurate customer data collection or updating in their rush to drive revenues. The high volume of data at financial institutions creates a very real risk that inaccurate data gets accepted without review. For service providers generating thousands of documents for mailing every day, the temptation can be to print and ship documents without verifying the accuracy of the data in data fields. This has the potential to lead to significant inaccuracies that could be interpreted as deceptive by regulators. Cyber fraud that uses stolen customer data and identity theft can create inaccurate customer data.

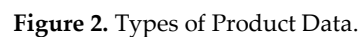
Disparate systems create unique challenges for customer data accuracy. For financial institutions of significant scale, mainframe back-end systems developed over decades are still in use. These systems were developed in different eras with different technologies, sometimes developed by different parties, or even different vendor versions. A significant number of these institutions have never invested in the effort to consolidate and modernize their customer data infrastructure and solution offerings, creating information silos that consistently fail to share onboarding data. These silos can also be created with new acquisitions that interface with the parent financial institution's systems. Without regular updating, address data may be continuously inaccurate. These mainframe systems were developed simultaneously as U.S. postal services were beginning the data-matching process, with the formation of service partners to validate and maintain post office data on the physical addresses used in mailings.

4. Product Data Integration: An Overview

Advancements in product data, and its integration, benefits financial services in a variety of ways. Product data management holistic platforms will help lower existing costs associated with disparate internal, non-centralized solutions. Enabling collaborative consumption approaches creates more enhanced, new services and financially liberating experiences built upon these product services and other basis services and enablers. And product data normalization and additional technology infrastructure simplifies integration of new partnership services taking days, not months. Foundation product service data enable predictive service algorithms that optimize customer service recommendation flows, opening up significantly to additional product features of vertically integrating to the overall shopping and post-use experiences. Essentially, shifting towards an omnichannel, continuous lifestyle marketing approach.

4.1. Types of Product Data

Product data covering naming information for every product should be uniquely identified. Product data should also include key-dimensional information, such as size or weight, and descriptive information, such as the 'about this item' information displayed on retailers' product detail pages. Descriptive information is required for lenders to optimize search, navigation, and merchandising elements for each product being offered. Lenders also need product data to provide product match lists for search queries on retailers' sites, display complimentary or alternative products on product detail pages, and provide visual identifiers, such as color swatches and product images, on product browse pages. In a similar way, retailers also must presume that lenders presently have product data that includes complete, recent, and accurate information on their entire product offerings.



Product Data Integration (PDI) is the practice of consuming, fusing, integrating, and aggregating data across disparate data sources. Product data from disparate sources aids brands in understanding the totality of the relevant product landscape. Product data can impact many business functions, including discovering the right product with the right attributes, shaping marketing using the right audience-defined textual semantics, addressing the right audience with relevant advertising, optimizing inventory mix and ensuring the right produce is available at the right time to address demand, undertaking pricing and offer management, and improving and customizing product design. Yet, like any other data, the full power of product data can only be leveraged when the data is integrated properly. Product data is inherently difficult to integrate because of the variability in product data semantics, including inconsistent attribute names, attribute

relationships and hierarchies, data format and representation, and so on; the diversity in possible data sources, including brand sites, retailers, comparison engines, image data, user-generated content, etc.; and the differences in update frequency and latency of the data sources. Today the consumer is boundless. As consumer behavior becomes more unique, the challenge of satisfying that behavior with the right product increases [8].

Product data integration is the action of bringing together all of the product information needed to analyze and discover insights about a product or category. It's used primarily by retailers, brands, and product agencies to ensure that the information stores throughout the enterprise are accurate and trustworthy. PDI is essential for bringing together product data from multiple internal and external data sources to create a single source of truth about a product or category. But creating that single source is often time-consuming and manually intensive, taking months or longer to complete. Without quality PDI, companies are unlikely to have high-quality assigned marketing for both branded and non-branded keywords. Search problems will impact launch and promotional timing and effectiveness, media spend and ROI, and product discovery. Without sound PDI as a foundation, companies may miss opportunities, and reviews and sales on other sites may significantly impact the company's own performance metrics. Because of these difficulties, enterprises typically succumb to working with small subsets of available product data sources. In recent years, developments in AI and machine learning have significantly helped address the challenges associated with PDI.

Equation 2: Master Data Quality Index (MDQI).

$$MDQI = \frac{C_{completeness} + A_{accuracy} + U_{uniqueness} + T_{timeliness}}{4}$$

Where:

- $C_{completeness}$: % of records with all mandatory fields filled
- $A_{accuracy}$: % of records without known errors
- $U_{uniqueness}$: % of unique customer/product identifiers
- $T_{timeliness}$: % of data updated within defined SLA

5. MDM Framework for Data Integration

In doing so, it represents a holistic view of the data integration processes with the unique focus on aligning data management with the MDM process and revealing the need for a framework that extends beyond the traditional boundaries of integrating data from a technical standpoint. The aim is to position our research contribution within this wider domain of data quality management work and connect the concept and goals of physical data integration with the processes and tools of MDM. Our conclusions lead us to suggest how physical data integration, and tools that assist with implementing it, might be integrated with the processes and tools of MDM to both gain efficiencies and, more importantly, better connect the data integration technologies to the business goals of an organization. Furthermore, enterprise data integration is a highly entrepreneurial and innovative technology space, and thus including this type of design and development space-based research is an important component of a well-rounded portfolio of research topics.

Data integration in the MDM sense is the responsibility of MDM because, as the strategic owner of the master domain, MDM determines what master entities and attributes are signed as trusted by the enterprise with the usual team of people behind it involved in the MDM workflows. MDM admits that physical data integration is a long-term ongoing thing and that part of the role of MDM is to enable physical data integration for those ones that provide master entities to the MDM hub what we would call trusted source systems and using the usefulness of a master domain to facilitate the normalizing

of the master space from the physical data integration standpoint for those missed master entities that still reside in operational systems typically not the trusted source systems.

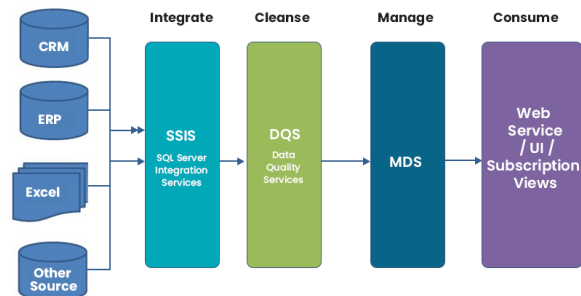


Figure 3. MDM Framework for Data Integration.

5.1. Key Components of MDM

Master data management (MDM) serves as a critical framework for data integration across services. MDM compiles, harmonizes, and curates information about the key entities around which financial services operate. MDM allows different services to share and consume the same data in an easy and consistent manner. It aggregates reference data such as country lists, address formats, currency codes, or types of transactions. This reference data often comes from different ranges of systems with different sets of logic. By establishing an MDM system that provides a clear guide of how reference data should look, services can easily discern incorrect reference data and either fix it or have a logical prevention to enforce the correct data entry. Such validation could steer people away from making wrong decisions. For example, address validation will reduce transaction errors such as undeliverable addresses and returned checks.

MDM implements a clear definition of the key entities to ensure consistency across services, regulations, and time. It stores core data such as users, accounts, and transactions. It publishes the data in an understandable manner, with different publishing response keys to suit the needs of various publishing services. Since every service has its own design guide, sometimes the same data is presented differently and in odd places. It is not the consumption of storage that is a concern, but the confusion that such discrepancies generate. With the fine granularity of the MDM system, each service can look to the MDM system for, say, user addresses and leverage that instead of creating and storing their own copies. The benefits of data validation, certainty, and consistency would lie with data in and out of MDM around the same entity. The other control in the design is the update lifecycle [9].

5.2. MDM Implementation Strategies

The implementation of a detailed and rich MDM layer for data integration in a federated model, apart from the typical ETL effort, needs guidance, enhancements and sometimes frameworks to ease the implementation effort and get to production quickly. Depending on the use case, the various stakeholders who will govern or access or consume the data, the current state of systems and future state use cases, there can be various strategies for implementing and supporting the MDM.

One specific strategy for a smaller number of entities with richer semantics is to hand-build them. This is similar in some ways to the design of a data warehouse, where complicated SQL and significant ETL effort is required to create the data. For the MDM project, the additional burden is that MDM hierarchies and PKs are something new — something that has to be built separately from any existing database design — and throughout some agreements have to be made to get PKs done within a tolerable assignment time. If some consortium of database designers from peer companies can be

assembled, and if joins can be inserted as potential PKs during the data loading, this hand-built method becomes more feasible. But experience has shown that it is usually quite slow. Most importantly, if the object-based way of embedding semantic MDMPKs and hierarchies into the operational GUI and transaction processes is in place, it becomes easier to build the initial MDM databases using the hand-built method.

Another strategy for the MDM process is to be “operational defaults.” Many business applications today can retain significant amounts of information within their schemas about the entities being managed. For example, customer files often have attributes such as “Account Type” or “Application Type,” and other attributes may specify the application among bank, credit, and investment. Extraction mapping can take advantage of this keyword information to segregate the entities. This introduces the notion of TCP objects; in effect, both the “tagging” of a data item as a key component of an entity and its attribute-rich data-ness are achieved using synchronizing the operational systems [10].

6. Data Governance in Financial Services

In financial services, trust in stakeholders involves confidence that organizational data assets are secure, accurate, reliable, complete, and verifiable. This trust can only be achieved through the implementation of proper data governance. Data governance is a function that includes processes, roles, responsibility, and decision rights to ensure data, and the data systems supported and enabled by it are properly managed throughout the lifecycle of data, from creation through asset retirement. Data governance ensures that business objectives are defined, and rules and standards are maintained so that the data adhere to these principles.

Data governance is both a top-down and bottom-up effort, as it combines strategic decision rights and the articulation of business principles about data with the tactically driven development of data standards and rules. It is important to have a top-down and bottom-up framework, as requiring all data governance processes to stem from the top-down approach is critical, as this approach supplies the necessary power and authority to ensure compliance by all staff, including IT. However, the continuous input and participation of the functional areas will be responsible for the success of the governance effort over the long haul. The bottom-up approach provides corporate technocrats with actual business line experience during the development and implementation phases. Within the data governance framework, top-down strategy flows into the business functions and back out through the IT function to shape system design, development, optimization, and stewardship. From the bottom-up perspective, business functions inform governance purposes and objectives and inspire applicable policies and standards [11].



Figure 4. Data Governance in Financial Services.

6.1. Establishing Data Governance Framework

Governance is the internal management oversight, reporting hierarchy, and definition of responsibilities to ensure organizational objectives are met. Data and technology must be at the forefront of overall company governance to ensure that as products scale and appropriate technology solutions are adopted, they are in line with the goals of the organization to ensure success at all levels. Once established, a data product and technology management function should meet with senior executives regularly to ensure that there is continual alignment among objectives, senior management focus, and progress of the product and technology roadmap and its priorities. Governance encompasses both the people and processes that work to ensure that data products are correctly building, maintaining, and operating to meet business objectives. Governance defines how, with whom, and when choices are made that affect the data product, technology, future roadmaps, and their integration with other products. It is important that awareness of these choices is pervasive throughout the organization. Data products impact business objectives but may span different areas within a company. It is therefore essential to provide a structure within which others can see a data product's strategy and vision, support it wherever possible, and understand that they have a subsidiary function in controlling the impact of their local decisions. At a team level, people within a company have their primary accountability for delivery. It is important to communicate that each team member owns the data products and that choosing to not create a data product or not commit to a release date, and data quality is not an acceptable option. It is why ownership is typically at a team level and does not devolve to a single product or tech person.

6.2. Best Practices in Data Governance

Data governance is a particularly challenging area given the flexible yet effective organization of the actual work. Business units have a deep knowledge of the area, and partnerships have provided support in many confirmed cases of specialization in detail and analysis needed. Data owners are typically responsible for whether it is the business unit, the data engineering team, the information technology that supplies the tools for data collection and processing, or a mix of all of those different players together. Data validation is to a large extent left up to the data owners and data submitters. Regulatory agencies stress that companies should have adequate disclosures around all data. In public company filings, that information is often buried, and used only to double check how "thick" the main corporate chapter is for financial reporting versus how "thin" other operations data chapters are for non-financial reporting operations. Poor disclosures around data quality may suggest potential issues.

Management/ownership of the same data subject/term across different uses and processes, with the same data dictionary entry, is critical for data quality, particularly if it has been centralized into a data warehouse by the IT function. That management is typically by business process area or system of record for the data, although for very large companies, this approach can grow convoluted. Concerns around data quality must also have strong internal auditor support. They serve as validators of the data quality statements provided, and can do enough analytical work to obtain a general feel for the validity of the data.

7. Data Quality and its Impact

Data quality refers to the condition of a dataset that is suitable for a particular purpose. While it is possible for datasets to be merely 'good enough', collective experience indicates that data quality does have an impact on downstream operations. Poor data quality can manifest as increased costs, wasted resources, missed opportunities, and more. Other effects within the financial services domain can include regulatory sanctions, ensuring operational resiliency, and increased fraud risk. In this section, we review the general interpretation of data quality before assessing how these measures tie into the

financial services domain. We then review the most commonly seen financial services-based data quality issues [12].

A practical framework is provided for measuring data quality, along with a short example to showcase the proposed approach. A healthy reluctance to oversimplify any topic should be maintained; as such, the following will hopefully not be misconstrued as the de facto list of considerations within data quality assessing frameworks. While data assessment is more often than not a subjective exercise, or at least tied to the purpose of how the data itself is intended to be used, there are nonetheless some major contributors to data quality that have become commonplace within the data engineering industry. The five main attributes empathized generally include: accuracy, comprehensiveness, consistency, relevancy, and timeliness of the datasets in question.

7.1. Defining Data Quality

Data has always been an important component of financial services decision-making, but only in the last two decades were financial services able to capture and store the vast amounts of data available around their activity. It is only now, with the resources available through cloud computing and the tools for access, transformation, and analysis, that organizations have been able to turn this huge resource into actionable insight. For this insight to be reliable and secure, data has to be of reliable quality. In recent years, with the focus on data valorization through well-governed and implemented processes, organizations have started to pay more attention to data quality as an organized and repeatable process. This process is demystifying the "magic" inside the data pipe, reducing the risk and uncertainty for the information products. Finally, organizations are internalizing the key message that "data is an asset" and that attention needs to be given to it, just as other asset owners do for their assets. In this chapter, we review the concept of data quality and describe the most important dimensions of data quality and how to measure them [13].

After having clarity on the concept of data quality and its dimensions, we then focus on data quality assessment tools and data quality monitoring, offering our perspective on best practices for assessment and monitoring. Because data products are the basis for information products, and because bad input data will lead to bad output information, decision-making organizations must invest in programs and tools to validate the data that is present in their data lakes or warehouses, at the point where they are either cleaned or integrated before being used by the analysis and data science teams. When answerable questions are being addressed in reports, dashboards, or data science models, and where the decisions resulting from such reports are key to the functioning of the organization, monitoring should ensure that the input and output data history is analyzed for inputs that fall outside acceptable levels or expected business events.

7.2. Measuring Data Quality

While both business and technical data quality assessments have metrics associated with them, business metrics are intentionally vague. Often supported by anecdotal evidence, the business measures tend to focus on outcomes and their association with success or failure. Technical measures on the other hand are often based on tangible computations such as records missing an entry in a specific field or records with malformed entries. Technical measures can be defined in reference to the intended use of the data product, and this common reference avoids the vagueness present in business measures, which also often reflect influences from different dimensions [14].

A major challenge of technical measures is that they are reliant on specificity. To be effective, a data product must be targeting exactly the use-case for which that measure was defined; otherwise, the results can be misleading. For example, consider two measures of data accuracy, one based on a validator sample of income values and the other based on knowledge that income values are correct for a different subset of records.

Both indicators can yield accuracy, but this information is not comparable and does not allow the user to assess which report indicates most accurate data.

In many cases, it is up to data product owners to ensure that technical measures and metrics of accuracy, consistency and completeness among others are monitored throughout the data product pipeline. A neglected step in creating data products is the scale at which an individual metric or corresponding service-level objective should be monitored. Events that occur with greater frequency – such as the presence of missing attribute values – are easy to expose at minor infractions. Lower-level gauges are typically reserved for rare conditional effects, like the distribution of values for a given attribute for data records associated with a local store. Compounding the issue of specificity is the level of dimensional decomposability, i.e. the ease or difficulty with which different measures of quality can be disaggregated. For example, the dimensional decomposition of value distributions by geography and time are reversible if both are included in the static feature distribution for addresses with a time component [15].

8. Integration Techniques for MDM

To build an operational view of foundational data, we need to extract data from multiple sources and disparate databases. Transform data, deduplicating, cleansing, and enriching it in some way, before loading it in a staging area in the operational database, or Master Data Management. Of course, that's the role of ETL processes. Extract characteristics in ETL processes focus on the variety and scale problems, as well as the need to clean and indicate unique keys before loading data into the operational database. Those tasks can be challenging in the operational data store of an MDM database. Data comes from different environments and can be on an operational data store with batch modes. The same data sources are queried numerous times and the time to extract data can be significant. The need for masking sensitive data increases the complexity.

The transformation and loading characteristics focus on the synchronizing and schema difficulties and loads either at the staging or incremental level after the bulk load is completed. The data can be loaded either from the ETL processes that extract to a staging area or on the fly through synchronizing, shimming, or replication modes. Actual loading and transformation can be done straight from source data during extraction. The source data can also supply temporary reference data to be used for the actual operations in the MDM, whether remote or loaded from a staging area.

For a Master Data Management database to maintain an accurate operational view, the back-end process must push individual records into the Master Data Management database in real-time. Such scratchpad databases must operate effectively within the transaction/operation processing systems and not at the reporting tool level. Set identification must not be done at the reporting level as is the case at the on-line analytical processing database role. Only high-level pseudo IDs should be passed in the Report by Operation Mode, reporting file or by external IDs.

8.1. ETL Processes

Master data management (MDM) solutions, including financial services MDM solutions, can often support various methods of integrating data from a variety of sources. The choice of integration method depends on several factors, including the capabilities of the solution being used, the service level requirements associated with the master data, the responsiveness requirements of all of the design and business processes associated with the master data, and the capabilities of the sources and other applications that depend on the master data. In this section, integration techniques are discussed in terms of the required speed and timing of data flows.

Historically, the most commonly used integration techniques are ETL (extract, transform, and load) processes. The ETL technique is strong on flexibility and high degree of transformation capabilities. ETL tools are sometimes used to move data from source

systems to analytic databases for business intelligence and data science workloads. In the case of MDM systems, subsequent loads may occur at frequencies designed to improve the freshness of master data associated with each data service level.



Figure 5. ETL Processes.

Recent years have seen the emergence of a number of vendor products and platforms that offer high-performance, simple, flexible, and nondisruptive ways to perform data extracts under almost any conditions imaginable. Because of its visibility and the issue of getting data into a "published" state, the official repository has a special requirement for loaded data quality. The specialized ETL offerings serve a variety of different use cases, including those from master data repositories, but data integration is central to MDM services, so most MDM solutions supply standard ETL capabilities [16].

8.2. Real-time Data Integration

Real-time data integration is a new capability becoming far more common and, in some cases, expected as data systems become more interconnected. One of the earliest forms of real-time integration was event-driven integration using middleware or similar technologies. In general, there are a few methods for real-time data integration. The oldest is the use of triggers on the source system to catch data changes as they happen and to send data event messages out to a broker to publish. This tends to be accounting/control-data-centric because it's mainly the transactional data that is changing and that is monitored. Some non-operational sources will publish status changes, too. The other common integration method is change data capture, where the source system has a changed data structure and can detect or store changes that occur and periodically export a snapshot of only changed records for external systems to import. Either can be used in combination with micro-batch processing to reduce latency.

For non-OLTP systems, there are message-publishing agents available to monitor for file changes or database changes. In addition, near real-time integration can be accomplished using either ETL or data replication technology with near real-time scheduling to perform short snapshot extracts and copies/loads. What's different from data update cycles typically seen with data mart ETL cycles or staging-area loads for central or replicated long-term data warehouse storage is that monitoring real-time data movement is not possible. Data movement with automated real-time integration just happens concurrently as data changes occur, with external monitoring for catch-up expected for correction [17].

9. Case Studies of Successful MDM Implementations

Although there is a wealth of information available on MDM, we were unable to find any publicly available descriptions of MDM implementations and their results. Consequently, we describe a couple of case studies based on assignments that the authors have worked on. We believe these cases will provide you with a much clearer picture of what an MDM implementation entails and the potential benefits of MDM. To the best of our knowledge, while there are a few companies that have published their case studies, they generally focus on how they selected and implemented the tool but not on how they actually built the data product after the initial implementation [18].

A large bank in the United States had purchased a multibillion-dollar trading platform to support its corporate client's foreign cash management and exchange services business. They had spent millions of dollars in customization and the platform was going live in less than six months. The application was directly integrated with the stock exchange, but the bank was concerned that the exchange would suddenly make it mandatory that all the bank's traded commodity and currency counterparties be validated with up-to-date credit ratings at the exchange and insufficient data hygiene could threaten business. The risk management team realized that the bank had no reliable data on its ecosystem of counterparties. Several business units had their own versions of counterparties, built over time through various activities such as KYC, onboarding, credit assessment, and loan origination, which were characterized by internal and external data silos and a poor data sharing culture. Each version had discrepancies such as different spellings of names or some names missing from other versions.

There was no way to know which names were missing or erroneously included in the various versions making them fallible for risk computations. The risk team surmised that the data travel time distribution for actual probable counterparties could be very skewed and with long tails thus increasing the possibility of crimes like money laundering. They needed a reliable enterprise-wide data product that had the all-time most updated likely copy pair of counterparties, so they could work with the respective business units to validate them.

9.1. Case Study 1: Bank A

In 2011, after over a decade of MDM project failures, an agile-minded top executive at Bank A mandated that the company establish a set of its master data domains and actually utilize them in a near / new real-time basis. Though the company had individual databases for each of its presumed domains, they were not connected, thereby presenting customer-facing employees and executives with varying and often contradictory versions of the truth. Since then, the MDM functions developed offered a single repository for ultimate sources of all data typically required for customer transactions. Easy access to the authoritative, pre-verified data pays huge returns through quicker and less costly resolutions of inquiries; and banking operations are decidedly less prone to error since all transactions utilized the same data source. Additionally, Bank A was able to automate much of the processing for EU Risk program by developing standard logic utilized in the underlying systems [16].

Unique business processes involving different approaches for each of its business lines and a corporate culture marked by a resistance to any centralization or standardization effort, Cyber Banking focused on the on-demand availability, versatility and reliability of the data. Bank A's vision for its MDM solution was the creation of a "consumption" structure to support the banks' significant Enterprise data Warehouse and transaction-based application efforts. Bank A MDM "solution" provided an easily understood set of services, using the underlying MDM System, the Enterprise Data Warehouse and transaction-based applications to establish a data management environment where the Data Consumers, both business lines and Corporate support functions, could initiate requests for the quality and availability of any data element utilized by them.

Equation 3: Data Product Value Score (DPVS).

$$DPVS = (V_{reuse} + V_{decision}) \cdot Q_{data} \cdot R_{compliance}$$

Where:

- V_{reuse} : Reusability value (how often data is leveraged across systems)
- $V_{decision}$: Value from data-enabled business decisions

- Q_{data} : Overall data quality score (e.g., from MDQI)
- $R_{compliance}$: Regulatory compliance factor (0-1 scale)

9.2. Case Study 2: Insurance Company B

Insurance Company B is one of the largest Vietnamese enterprises in life insurance and financial investment. Their vision is to become a pioneer in providing a comprehensive system of financial and investment solutions to people on the way to increase their wealth, build assets, and realize their noble dreams. Similar to Bank A, the emphasis of guiding policies on developing e-business channels is required and planned for long in Insurance Company B operations. In addition, the establishment of a legal framework, regulations, and business processes for online insurance services is being completed. Insurance Company B's target is 10% of the total insurance market share. The share from e-business channels should reach 7%, by then earning premiums and reflecting profits, contributing the most amount to the insurance industry compared to other industries [19].

The Investment Project for Building Infrastructure for Online Transactions in Insurance Business is undertaken and being implemented by Information Technology Company B. Accordingly, Insurance Company B uses its own technical foundation and operates an online insurance agent system and online customer management. The implementation aims to generalize insurance data throughout the country, serve customers completely, and utilize insurance data to develop and maintain long-term relationships with customers and sales agents. The results after nine months of implementation show that the gradual investment development and explosion of the online insurance market will positively impact and benefit the entire economy. By early 2001, Insurance Company B had achieved the second position in terms of customer service on the Internet. By the end of 2001, Insurance Company B expected to conduct a sizable telecommunications operation, while its transaction volume had exceeded the inter-operational market.

10. Future Directions in Data Product Development

The chapter summarizes a wide range of data product development topics covered in the book. The final chapter explores practical innovations in various stages of data product development. It starts with describing automation enabled by platforms that impact the upstream stages of data product lifecycle such as data integration and modeling, and continues with innovations in versioning and lifecycle management for data products. We then cover automation and AI/ML-driven innovations in the downstream data product stages of Data Analytics. Finally, we discuss the role of data integration and management within organizations and the external data and resource sharing across organizational boundaries [17].

While the foundational systems created within data products and services today do "heavy lifting" for companies, the specialist teams create data that is aligned with the business and responsible for less than 5% of costs. Thus, continued innovation in the foundational systems available for data products can move this ratio even further - enabling teams to deliver more value on a smaller budget. We identified a few areas in need of innovation, such as better integration environments supporting faster delivery time and adaptable infrastructure. This would help bring the creation of these components closer to the business stakeholders. Expedited operations are also needed in the area of Data Analytics such as self-service reporting to focus more on analytics and insights enrichment and less on reporting, especially in financial services, where reporting poses an ever-growing challenge.

Another area of innovation and discovery is the connection between technologies and data products. This connects data of a business function/process with the internal

activities and external parties, thus enabling shared treatment of data in multi-party infrastructures or services being efficient, of high quality, and accessible on demand. Thus, it helps create a new layer of reusable data services across company boundaries and vertically across supply chains, from various product sources to the consumers, all of which can be leveraged in the Data Analytics step of data products across domains [20].

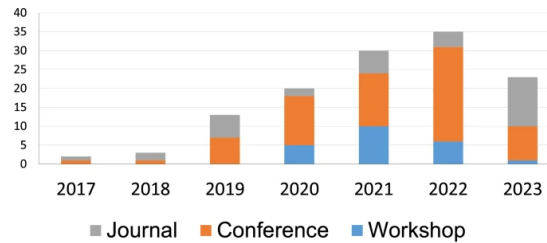


Figure 6. Building Foundational Data Products for Financial Services

10.1. Innovations in Data Integration

The next major wave of data product development will focus on the next bottleneck in the cycle of data product development: integrating disparate data products into product solutions. Next to actual domain knowledge regarding the particular use case, a critical success factor in developing data products is the ability to integrate multiple data sources efficiently. This requires specialists who often build data pipelines that copy data from multiple products into data warehouses. The step is tedious, expensive in time, resources and money, and does not add to the actual analytical value of a data product. It is therefore not surprising that many data products are less used than they should be. The research community informally refers to this step as a data spaghetti, a plethora of proprietary interfaces that make integrating data products a huge effort [21].

In the future, innovative integrators of data products will emerge that focus on solving the data spaghetti. In the domain of financial services, we already see signs of this integration wave. Data warehousing solutions that offer powerful SQL interfaces for all business users will receive a serious overhaul. In a new data warehousing generation, they will allow data analysts to work directly with data products across multiple data product sources without copying the data into their own home data warehouse.

While the new solutions will be able to exploit and accelerate the SQL query if possible, they will also create an agile environment for demand-driven data product integration where no data analyst will be slowed down by any physical data model design. Providing data products with specific drivers that allow integrated access to all data product capabilities will be as common as providing commodity drivers for databases today. Integrators will not own the data; their databases will act as middleware and routing tools for the distributed and federated data product networks [22].

10.2. The Role of Blockchain in Data Management

Blockchain technologies can have a considerable impact on the science of data management. On one hand, the ongoing innovation in the core of the blockchain ecosystem can create challenges for the rest of the data management ecosystem. On the other hand, the innovative components of the blockchain technology portfolio can also signal new directions for the future of data management. This chapter describes these different perspectives.

The guarantees that blockchain systems are meant to offer are often considered to be good enough to replace more general and richer data management systems. The lack of support for data queries and complicated transaction languages is sometimes viewed as a strength rather than a weakness, as it ensures that blockchain implementation will be immune to security problems that plague larger systems. However, this also means that

if you choose to model your data as blockchain objects, you give up capabilities that are being offered to you by general data engineering and management systems. The main perspective is that for certain characteristics of the dataset, answers to queries can be certified without a traditional query engine. In particular, this can sometimes be done with sufficient efficiency that dedicating a separate layer to the validation of answers to queries, without the support of a query system, is convenient.

Data Certification is an idea that is old in the field of data quality, and where the certification has classically relied on the judicious selection of "expert" users who would vouch for the accuracy of the data that they created. The earliest works on the idea described signatures from trusted peers, much like a digital signature for a piece of data, and its application to building a market of certified data. Although original, these ideas were clumsy in practice. The cryptographic primitives proposed were not effective, and the different amount of work on creating the piece of data created a financial asymmetry that would make the market impractical. A solution where the certification is based on folksonomies rather than natural hierarchies was thus proposed [23].

11. Conclusion

A review of the thematic areas addressed throughout this work demonstrates that the process of ideation and solution delivery across financial service firms presents opportunities to leverage market knowledge, technology, design, and analytics to innovate and create new value in areas of user experience, product enhancement, and cost reduction. These principles hold true to both point solutions to existing business problems and foundational data methods which accelerate the development and deployment of data insights across these organizations. It is through an execution as described in this work, framed against the needs of foundational data, that we can elevate the practice of building such data products into inversions of traditional software development activity. With examples across large and small organizations that have successfully integrated the principles, we present insights and considerations for transforming innovation practice into a co-creating capability for authors internal to the organization, and partners who seek to support successful product activity with the setup of the environment for success. The result is a company that is fully enabled in the expression of solutions that differentiate them in an increasingly challenging market environment, where customer loyalty is a rare and precious corporate attribute. We sincerely hope that this work serves as a catalyst for such transformations, and presents even a portion of its authors such satisfaction in verse, that they endeavor to pen their own story.

References

- [1] Kalisetty, S. (2017). Dynamic Edge-Cloud Orchestration for Predictive Supply Chain Resilience in Smart Retail Manufacturing Networks. *Global Research Development (GRD)* ISSN: 2455-5703, 2(12).
- [2] Vamsee Pamisetty. (2020). Optimizing Tax Compliance and Fraud Prevention through Intelligent Systems: The Role of Technology in Public Finance Innovation. *International Journal on Recent and Innovation Trends in Computing and Communication*, 8(12), 111–127. Retrieved from <https://ijritcc.org/index.php/ijritcc/article/view/11582>
- [3] Mashetty, S. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. *Analytics, And Innovation* (November 30, 2020).
- [4] Connected Mobility Platforms and Their Impact on Urban Transport Planning and Service Design. (2019). *IJIREEICE*, 7(12). <https://doi.org/10.17148/ijireeice.2019.71208>
- [5] DevOps Enablement in Legacy Insurance Infrastructure for Agile Policy and Claims Deployment. (2019). *IJIREEICE*, 7(12). <https://doi.org/10.17148/ijireeice.2019.71209>
- [6] Chakilam, C., Koppolu, H. K. R., Chava, K. C., & Suura, S. R. (2020). Integrating Big Data and AI in Cloud-Based Healthcare Systems for Enhanced Patient Care and Disease Management. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12), 19-42.
- [7] Pandiri, L., Singireddy, S., & Adusupalli, B. (2020). Digital Transformation of Underwriting Processes through Automation and Data Integration. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12), 226-242.

-
- [8] Lakkarasu, P. (2020). Scalable AI Infrastructure: Architecting Cloud-Native Systems for Intelligent Workloads. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12), 133-151.
 - [9] Big Data and Machine Learning in Fraud Detection for Public Sector Financial Systems. (2020). *IJARCCCE*, 9(12). <https://doi.org/10.17148/ijarccce.2020.91221>
 - [10] Yellanki, S. K. (2016). Smart Services and Network Infrastructure: Building Seamless Digital Ecosystems. *Global Research Development (GRD)* ISSN: 2455-5703, 1(12), 1-23.
 - [11] Botlagunta, P. N., & Sheelam, G. K. (2020). Data-Driven Design and Validation Techniques in Advanced Chip Engineering. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12), 243-260.
 - [12] Somu, B. (2019). Autonomous Agent-Based Systems for Real-Time Credit Risk Assessment and Decisioning. *Global Research Development (GRD)* ISSN: 2455-5703, 4(12), 46-69.
 - [13] Meda, R. (2019). Machine Learning Models for Quality Prediction and Compliance in Paint Manufacturing Operations. *International Journal of Engineering and Computer Science*, 8(12), 24993–24911. <https://doi.org/10.18535/ijecs.v8i12.4445>
 - [14] Pamisetty, V. (2020). Optimizing Tax Compliance and Fraud Prevention through Intelligent Systems: The Role of Technology in Public Finance Innovation. Available at SSRN 5250796.
 - [15] Pandiri, L. (2019). Leveraging Deep Learning for Automated Damage Assessment in Condo and RV Insurance. *Global Research Development (GRD)* ISSN: 2455-5703, 4(12).
 - [16] Kummari, D. N. (2020). Machine Learning Applications in Regulatory Compliance Monitoring for Industrial Operations. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12), 75-95.
 - [17] Gadi, A. L. (2019). Enhancing Vehicle Lifecycle Management through Integrated Data Platforms and IoT Connectivity. *International Journal of Engineering and Computer Science*, 8(12), 24973–24992. <https://doi.org/10.18535/ijecs.v8i12.4443>
 - [18] Pamisetty, A. (2019). Big Data Engineering for Real-Time Inventory Optimization in Wholesale Distribution Networks. Available at SSRN 5267328.
 - [19] Yellanki, S. K. (2017). Service Integration in B2C Ecosystems: Enhancing Customer Journeys with Intelligent Automation. *Global Research Development (GRD)* ISSN: 2455-5703, 2(12).
 - [20] Meda, R. (2020). Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. *International Journal of Engineering and Computer Science*, 9(12), 25289–25303. <https://doi.org/10.18535/ijecs.v9i12.4587>
 - [21] Yellanki, S. K. (2019). The Future of Efficiency: Integrating Consumer Feedback Loops in Digital Platforms. *International Journal of Engineering and Computer Science*, 8(12), 24928–24946. <https://doi.org/10.18535/ijecs.v8i12.4447>
 - [22] Meda, R. (2020). Real-Time Data Pipelines for Demand Forecasting in Retail Paint Distribution Networks. *Global Research Development (GRD)* ISSN: 2455-5703, 5(12).
 - [23] Somu, Bharath. "Machine Learning for Predictive Maintenance in Banking Infrastructure Services: A Data-Centric Approach." *International Journal of Science and Research (IJSR)*, vol. 9, no. 12, 2020, pp. 1948-1957, <https://www.ijsr.net/getabstract.php?paperid=MS2012141601>, DOI: <https://www.doi.org/10.21275/MS2012141601>