Middesk

# Data quality and strategy for modern business compliance

# Contents

# Introduction

For regulated financial institutions, following AML regulations is crucial to ensure compliance and safeguard against financial crime. Sourcing and utilizing data about individuals and businesses plays an enormous role in the ability of these financial institutions to evaluate them for risks. Enabling swift onboarding of legitimate customers while keeping out bad actors at scale is a top priority and therefore the quality of data directly impacts key business outcomes, like auto-decisioning rates in business onboarding. While opinions vary on what constitutes "good" data quality, its importance in combating financial services fraud is undisputed.

Over two decades ago, there was a wide emphasis on the volume of data financial institutions could use for their decisioning. Large scale data aggregators that combine many sources of data over time and return them to financial institutions played a vital role in meeting Patriot Act requirements, especially in manual investigations and due diligence. But, as onboarding digitizes and online forms replace in-person applications, data needs have shifted dramatically. Large volumes of data provided by data aggregators are consistently old and duplicative and confusing. These data sets drive up the need for manual, staffed reviews to make decisions; but manual reviews of data cannot keep pace with the growth institutions want and the speed that customers expect. Yet, these data issues are exacerbated by the risks financial institutions face with more sophisticated bad actors and AI advancements. Efficient, high-quality decisions about risks are more critical than ever.

At the heart of these decisions lies the quality of the data being utilized. Decisions are only as reliable as the data they’re based on and foundation of flawed data compromises evaluation and system built on top of it. This is particularly true for Anti-Money Laundering (AML) practices, where the strength of a financial institution's defenses is directly tied to the integrity and quality of the data that powers their decision-making frameworks. Without high-quality data, even the most sophisticated models and processes risk failure.

At Middesk, we have a vision that high-quality data can solve these problems. By leveraging authoritative, alternative, and first-party data sources, constantly refreshing that data, and delivering it in real time to financial institutions, Middesk has built the most comprehensive business dataset and platform on the market. Our goal is to power fast, confident decisions that drive the global economy. And our customers enjoy the ability to confidently navigate compliance, prevent fraud, and drive efficient revenue growth.

High quality data is what we focus on each and every day. And we believe that understanding what high-quality data means is crucial for AML leaders to successfully design best in class programs.  What follows is our introduction to data quality to empower you to more effectively identify, leverage, measure, critique, and ultimately, improve the data quality used in your business.

# Data quality and financial institutions

## Why does quality matter?

At its core, high-quality data can enhance decision making, reduce risks, and improve customer satisfaction for financial institutions.

## Business fundamentals

According to 2023 U.S. census data, three businesses are formed for every two people born in the United States. This surge is largely attributed to entrepreneurs meeting pandemic-induced economic needs, with significant growth in retail and professional services. Furthermore, small and medium-sized businesses (SMBs) utilize financial services heavily. While comprising under 20% of payment flows, The U.S. SMB segment accounts for about 55% of potential revenues for acquirers.

Keenly aware of these trends, financial institutions see capturing SMBs as an immense lever for growth while under pressure to improve efficiency. Despite the enormous potential of this market segment, these SMBs pose logistical challenges. As the volume of businesses increases, managing and utilizing data effectively becomes increasingly complex. Overall, SMBs have lower automated pass rates through onboarding flows due to low quality data matching against their newly formed business records. The expectations of these buyers are also evolving. They now demand the speed, frictionless experience, and automation they enjoy as consumers. When they do not receive it, they abandon the financial institution. A recent poll found that 3 out of every 4 businesses have abandoned their onboarding process to financial services due to poor user experience. Furthermore, these clients are extremely loyal—making capturing them first paramount.

## From a regulatory perspective

There's an inherent tension between driving growth and managing increasingly frequent risk. A Liminal report highlights that 62% of large banks witnessed a surge in financial crime and fraud in 2023.

The magnitude of impact from onboarding a bad actor can be severe, leading to regulatory fines, reputational damage, and financial loss. In 2021, 98% of B2B firms reported fraud attacks, losing an average of 4% of their annual revenues. More than two-thirds of these firms expressed dissatisfaction with their current fraud prevention methods and cited a plan to improve their AML practices.

How are leaders in the financial services industry tackling this complex environment? **They are looking to better data as the solution.**

They are seeking the best quality, freshest data to fuel automated processes for comprehensive fraud prevention measures while maintaining a focus on enhancing the customer experience. A recent Liminal report emphasizes the importance of data quality as a key factor in purchasing decisions for business onboarding solutions. In fact, 88% of respondents identified outdated or inaccurate business entity verification data as a primary driver of complications, delays, and fraud. Outdated data increases the likelihood of fraud by inaccurately portraying businesses, or causes friction by necessitating additional manual reviews.

## Data Quality Mastery: Middesk's vision

At Middesk, we believe that well-run compliance and risk programs, built on a high-quality differentiated data platform, can become a powerful growth lever for a business. We believe high-quality data fuels automation, reduces risks, and enhances the customer experience, ultimately driving growth and efficiency in the business onboarding process. Our team is fanatical about data quality, and our philosophy and company strategy reflect that commitment.

# Breaking Down Data Quality

## Defining Data Quality

We inherently understand the concept of "quality" in many areas of life—whether it's quality of life, quality time, quality control, or the quality of sound, food, and air. These are terms we grasp intuitively. However, data quality is more abstract.

At Middesk, data quality is a key differentiator. Our solutions are built on a unique foundation of data, and the effectiveness of our offerings is directly tied to the quality of that data. When we talk about data quality, we're referring to our rigorous approach to selecting data sources—ensuring they are both reliable and comprehensive.

The importance of data quality has been recognized for over three decades in information science and technology, gaining even more prominence with the rise of computing and machine learning. In today's digital landscape, understanding and ensuring data quality is vital for building compliant and effective decision-making processes.

Ensuring high data quality isn't just about having accurate or complete information—it's about creating a foundation for making informed, timely, and cost-effective decisions. As businesses and regulatory demands evolve, the importance of data quality becomes clearer, underscoring its role in driving compliance and operational efficiency.

To further illustrate this, let's explore a few definitions of data quality that capture its essential characteristics.

**Consider these definitions of data quality:**

- Data quality is the probability of data being used effectively, economically, and rapidly to inform and evaluate decisions.
- Data quality is increasingly seen as a measure of the reliability, completeness, and accuracy of data as it relates to the real-world state it represents.

**Data Quality**

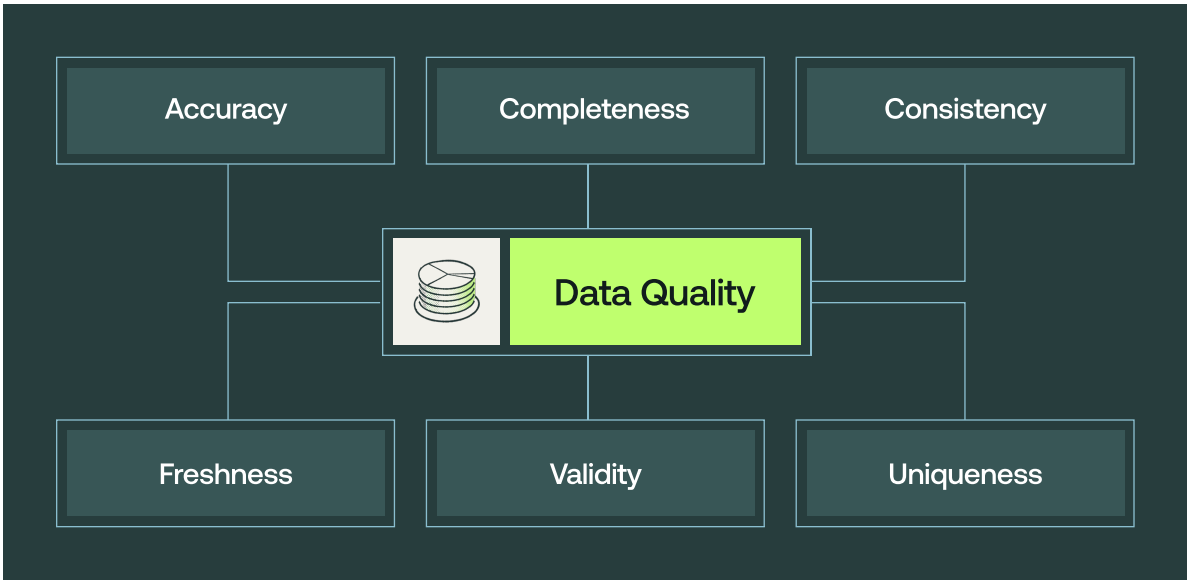
In simpler terms, high-quality data accurately reflects the real world, which is inherently complex and multi-dimensional. This complexity makes data quality a multi-faceted concept that requires careful consideration

To truly understand data quality, specificity is crucial. Just as we define "product quality" as a product consistently meeting customer expectations, we must define "data quality" in terms of its reliability, accuracy, and relevance. At Middesk, our products are built on a solid foundation of data, making the quality of that data fundamental to the effectiveness of our solutions. Therefore, clearly defining and understanding data quality is vital to delivering reliable, high-performing solutions that meet the stringent demands of our clients in financial services and other regulated industries.

## Dimensions of Data

Understanding the multi-dimensional nature of data quality is essential for organizations aiming to leverage data effectively. At Middesk, we define data quality through various dimensions, each representing a specific attribute that contributes to the overall quality of the data. What follows is our philosophy on how we shape data quality by six key dimensions:

1. Accuracy
2. Completeness
3. Consistency
4. Freshness (timeliness)
5. Validity
6. Uniqueness




Data quality concerns can arise at various levels within the data delivery chain, each contributing to the overall integrity of the information. You can conceptualize this as different levels - from very specific, small bits of information - called an “attribute” through increasingly larger levels of grouping that are multi-faceted collections of datasets. This flows from: attribute, record, dataset, product and customer levels of data. For the purposes of this discussion, we’ll mention attributes, records and data sets below. No matter what the level, data quality should be high in all six dimensions.

## Accuracy

Accuracy measures how closely data values align with the source of truth. Of the many dimensions, it's the most encompassing and arguably most loaded term – perfect accuracy is not only precise, but specifies that the data perfectly matches the agreed source.

It is often regarded as the quintessential marker of data quality by our clients, as it ensures that the data is precise and reliable, directly impacting their ability to make informed decisions, meet regulatory requirements, and mitigate risks effectively. A high level of trust in data accuracy is critical for maintaining the integrity of their operations.

Metrics such as error rate and auto-decision rate help gauge accuracy. We measure error rates based on the frequency of our customers reporting an inaccurate result to an attribute value not matching its source. The auto-decision rate reflects the trust customers place in automated decisions based on a combination of the workflows they have built and the high-quality data that powers those workflows.

99%

### Accuracy

Middesk offers **99% accuracy** because of proprietary data pipelines to 52 Secretary of States.

## Completeness

Completeness refers to the expectation that data is sufficiently provided to deliver meaningful inferences and decisions. Certain attributes may be expected to have assigned values in a record - or single entry within a dataset that may contain multiple attributes (or fields). For example, in a database, a record might represent a single customer, a single transaction, or any other entity being tracked. Certain records are expected to be present or not present in a dataset. We are defining data completeness as the data having:

1. mandatory attributes with values
2. optional attributes that may have values
3. and inapplicable attributes without values.

Our buyers rank completeness as a key marker for quality data because it directly influences the reliability and utility of the information they rely on. Metrics such as fill rate and qualification rate are essential in assessing the completeness of data, ensuring it meets their expectations and supports informed decision-making. For instance, fill rate measures the extent to which data attributes, such as business addresses, are fully populated in a dataset. This metric helps gauge the coverage of the data. Similarly, qualification rate evaluates whether data meets specific criteria, such as having a domestic registration record, to determine its completeness and relevance.

100%

### Coverage

Middesk offers complete, up-to-date information on 100% of registered businesses in the U.S.

## Consistency

Consistency ensures that information remains uniform and comparable across various sources and over time. When data values for the same entity or attribute are consistent, organizations can trust that their analyses, reports, and decision-making processes are based on accurate and reliable information. Inconsistent data, on the other hand, can lead to discrepancies and contradictions that undermine the integrity of business operations.

To maintain consistency, organizations often rely on metrics such as anomaly detection and manual inconsistency checks. Anomaly detection tools can automatically identify data points that deviate from expected patterns, flagging potential issues that require further investigation. Similarly, manual inconsistency checks involve comparing data across different sources, datasets, and time periods to identify and resolve discrepancies. By regularly monitoring these metrics, organizations can pinpoint areas where consistency may be lacking and take corrective action to ensure their data remains trustworthy.

+125

### Consistency

Middesk delivers a consistent data format and structure for a business, normalizing attributes and their values and aligning schemas across over 125 disparate datasets. We constantly evaluate and add new datasets to further enrich this data.

## Freshness (timeliness)

Freshness emphasizes the importance of data being available and ready for use precisely when it is needed. This concept extends beyond mere availability to encompass the recency of the data—whether it reflects the most up-to-date information from the source of truth. In fast-paced environments where real-time decision-making is critical, the timeliness of data directly impacts operational efficiency and the ability to respond to changing circumstances.

To measure and maintain this timeliness, organizations rely on metrics such as **data freshness** and **monitoring lag**. Data freshness assesses the "currency" or "recency" of data by measuring the time elapsed between when data becomes available from its original source and when it is accessible within the system. For example, tracking the time between the availability of registration data from various Secretaries of State and its incorporation into the dataset is a practical application of this metric. Monitoring lag, on the other hand, evaluates how closely a system can report on events in near real-time, capturing the speed at which changes are observed and reported. By regularly assessing these metrics, organizations can ensure that their data remains timely, enabling them to act swiftly and efficiently in response to the availability of new information.

92%

### Records refreshed in past 10 days

Middesk offers the freshest data available to drive insights for our customers.

## Validity

Validity ensures that data values are correct, meaningful, and relevant within their specific context. Valid data aligns with defined standards and guidelines, making it not only accurate but also applicable to the intended purpose. When data values are valid, organizations can trust that the information they are using for decision-making is both reliable and pertinent, which is crucial for maintaining the integrity of business processes and outcomes. Without this assurance, decisions based on inaccurate or irrelevant data can lead to misguided strategies and inefficiencies.

To assess and maintain validity, organizations often utilize metrics such as the invalid value count or rate. This metric measures the number or percentage of data values that fail to meet established data quality standards. For example, a field like "status" might be expected to contain only specific values such as "Active," "Inactive," or "Unknown," and any deviation from this set—or a null value—would be considered invalid. By monitoring these metrics, organizations can identify and address areas where data may not conform to required standards, ensuring that only valid, high-quality data is used in their operations.

160M

### Validity

Middesk cleans and validates over 160M business-related addresses, ensuring they are valid, linkable, and available for further enrichment.

## Uniqueness

Uniqueness indicates that each data record or entity is distinct and free from duplicates or redundant entries. Unique data guarantees that every data point represents a single, unique instance, preventing the complications that arise from multiple data entries for the same entity. Uniqueness is essential for maintaining data integrity, as it ensures that analyses and decisions are based on accurate, unduplicated information. Without this assurance, data could be misinterpreted, leading to errors and inefficiencies in business onboarding processes.

To measure and maintain uniqueness, organizations often rely on metrics such as duplication count or rate. This metric identifies the number of duplicate or redundant entries within a specific dataset or across multiple data sources. For example, when searching for a business in a database, the goal is to return a single, accurate record that reflects the business's current and real-world state. However, if the business is split across multiple records, this duplication could lead to confusion and inaccurate analysis. By monitoring duplication metrics, organizations can address and eliminate redundancies, ensuring that their data accurately represents unique instances, thereby enhancing the overall quality and reliability of their data.

170M

### Uniqueness

Middesk links and deduplicates over 170M business records to generate a unique profile for every business entity in the United States.

# What makes Middesk's data strategy unique?

Middesk's data strategy sets it apart in the business onboarding landscape by addressing the common frustrations compliance professionals face with alternative solutions. Often, these solutions rely on third-party data sources that fall short in accuracy, completeness, and relevance, leading to lower automation rates and longer turnaround times. At Middesk, we overcome these challenges by building direct pipelines to 51 Secretary of States and other high-quality data sources, ensuring our customers have access to reliable and comprehensive information. This approach empowers fast, confident decision-making, as our proprietary data pipelines cover 100% of registered businesses in the U.S., with 92% of records updated within the last 10 days. Additionally, we maintain established connections to other authoritative sources like the IRS, FMCSA, NPPES NPI, and federal and state courts.

Our unique data platform drives high-quality business onboarding decisions by focusing on areas where we have a distinct data advantage. By leveraging our expertise and comprehensive data access, we deliver meaningful insights that are unmatched in the industry. For example, our platform provides unique address risk insights and uncovers relationships between entities, enabling our customers to identify potential risks and connections that might go unnoticed with other vendors. This capability is crucial for making informed business decisions and enhancing the overall quality of the onboarding process.

Our vision is that every business has access to the information they need to make fast, confident decisions about how to operate in their market and serve their customers. To fulfill our vision, we continuously explore and add new datasets to our offerings. We currently have about 100 datasets in production, each representing a distinct data pipeline. Our ability to rapidly integrate new datasets is a key differentiator. This speed and efficiency are the result of significant improvements to our platform, including the automation of data ingestion, adoption of a flexible machine learning model, and a 300-fold increase in processing speed, enabling us to process 6,000 profiles per second.

Through these advancements, Middesk delivers a comprehensive and unified picture of each business, accessible instantly through our API and user-friendly dashboard. Our platform's ability to standardize and connect disparate data points into a single business identity ensures that our customers can make informed, data-driven decisions with ease and confidence.

## Conclusion

The words of W. Edwards Deming resonate profoundly: "Good quality means a predictable degree of uniformity and dependability with a quality standard suited to the customer." At Middesk, we recognize that superior data quality is the foundation upon which trust is built—trust in the data, the product, and the company itself. By consistently delivering high-quality, reliable data, we empower our customers to make informed decisions with confidence, fostering stronger business relationships and driving meaningful outcomes. As the landscape of business onboarding continues to evolve, maintaining this standard of excellence remains our unwavering commitment, ensuring that our clients can always rely on Middesk as a trusted partner in their success.


Middesk, Inc.