myridius

The Data Testing Imperative

How to Safeguard Business Success in a Digital World

AUTHOR

SOUBHAGYA PATRA
Myridius Director,
Data Testing Capability

NIKO MANGAHAS
Myridius Head of Quality
Engineering and Automation

INTRODUCTION

Why robust data testing is a business necessity

Enterprises across sectors are increasingly relying on data to drive informed decision-making, but their data quality isn't always keeping pace. This growing "data debt" crisis poses potential detrimental downstream impacts on business intelligence and operational effectiveness. Gartner reveals a sobering reality: Bad data costs enterprises an average of $12.9 million annually[1].

This is where data testing comes in as a critical safeguard. It is more than a technical function but serves as a strategic imperative that validates data across key dimensions: accuracy, consistency, completeness, reliability, and timeliness. Data testing identifies anomalies and errors early, helping teams to mitigate risks and prevent disruptions in data-driven operations. Beyond just preventing mistakes, data testing builds a foundation of trust, efficiency, and competitive advantage.

Modern businesses face a complex array of challenges: increasing administrative complexity, a shortage of skilled data professionals, escalating regulatory demands, and rapidly evolving customer expectations. Robust data testing provides a powerful solution that offers a systematic approach to maintaining data integrity in an increasingly unpredictable business environment.

Data testing has evolved significantly, from basic data integrity checks in traditional databases to sophisticated validation frameworks in modern cloud and AI-driven ecosystems. As businesses increasingly rely on real-time analytics, the need for robust data testing has grown beyond simple schema validation to include automated quality checks, anomaly detection, and business rule validation. This ebook explores best practices for ensuring accurate, reliable, and high-performing data pipelines.

1 - Gartner, "How to Improve Your Data Quality", July 2021

In prioritizing rigorous data testing, enterprises can:

- Mitigate financial risks associated with poor data quality
- Enhance operational efficiency and decision-making capabilities
- Build a culture of data reliability and transparency
- Enable accurate and trustworthy AI models with accurate and clean data
- Reduce the risk of regulatory non-compliance
- Stay agile in a rapidly changing technological landscape

In an era where data is the new currency, comprehensive data testing is not just a best practice—it's essential for sustained success.

The Risks and Consequences of Poor-Quality Data

Poor-quality data can lead to significant organizational challenges:

FINANCIAL LOSSES

Compromised data integrity can skew financial assessments, trigger revenue leakage, and generate billing inaccuracies.

REGULATORY AND COMPLIANCE VIOLATIONS

Data inaccuracies can result in compliance breaches, leading to hefty penalties and legal consequences.

ERODED CUSTOMER EXPERIENCE

Inaccurate customer information compromises service delivery, fragments communication strategies, and undermines marketing effectiveness.

OPERATIONAL DISRUPTIONS

Data inconsistencies create friction across business processes, causing supply chain interruptions, logistical delays, and systemic workflow impediments.

MISGUIDED BUSINESS DECISIONS

Unreliable or inconsistent information can lead to fundamentally flawed strategic decisions and misaligned organizational objectives.

Understanding Data Testing

Data testing is the systematic process of verifying and validating data throughout its lifecycle to ensure accuracy, completeness, consistency, timeliness, and validity. It goes beyond simple quality checks to serve as a comprehensive approach for maintaining data integrity in complex business environments. By identifying and mitigating discrepancies at key stages—ingestion, transformation, storage, and retrieval—organizations can ensure their data remains a reliable strategic asset.

Effective data testing involves multiple layers of validation, including:

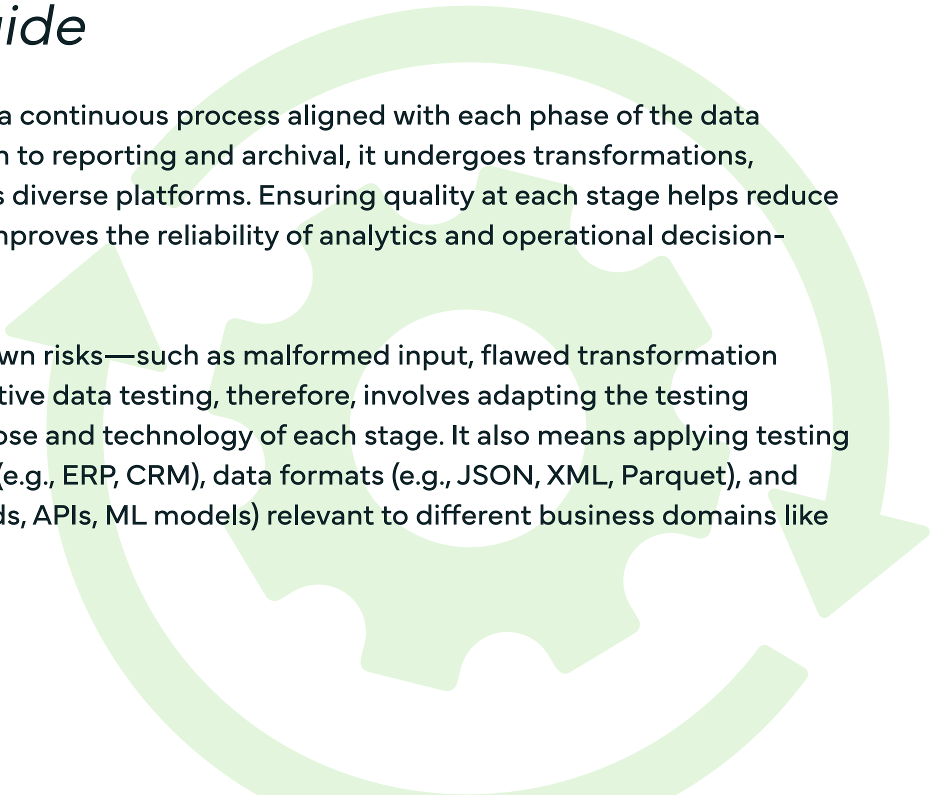
- **Validating raw data extraction** from source systems.
- **Verifying transformation logic** against business rules.
- **Ensuring reconciliation** between source and target data.
- **Monitoring data quality** in production environments.

This rigorous approach ensures that data pipelines function as intended, enabling trustworthy analytics, reporting, compliance, and AI/ML initiatives.

Testing Across the Data Lifecycle
A Comprehensive Guide

Data testing is not a one-time activity but a continuous process aligned with each phase of the data lifecycle. As data moves from raw ingestion to reporting and archival, it undergoes transformations, validations, enrichments, and usage across diverse platforms. Ensuring quality at each stage helps reduce downstream issues, promotes trust, and improves the reliability of analytics and operational decision-making.

Each stage of the lifecycle introduces its own risks—such as malformed input, flawed transformation logic, or outdated metrics in reports. Effective data testing, therefore, involves adapting the testing techniques and tooling to match the purpose and technology of each stage. It also means applying testing strategies that reflect the source systems (e.g., ERP, CRM), data formats (e.g., JSON, XML, Parquet), and the consumption methods (e.g., dashboards, APIs, ML models) relevant to different business domains like banking, insurance, and healthcare.



Data testing occurs at different phases of the data lifecycle. Each phase requires specific types of validation to ensure the data maintains its integrity, usability, and compliance. Testing needs also vary based on the source and type of data being ingested, transformed, stored, and consumed.

1. Data Source Phase

At the data source stage, testing ensures the integrity, structure, and readiness of data before it enters processing or transformation pipelines. Since data originates from diverse sources—including file-based systems (e.g., CSVs from external vendors), relational databases (e.g., Oracle, SQL Server, PostgreSQL), application APIs (e.g., CRM, ERP, eCommerce), and enterprise data platforms (e.g., SAP, Salesforce, Workday)—each source introduces unique risks and validation requirements. Poor data quality at this stage can lead to cascading errors throughout the pipeline, making early detection essential to maintaining accuracy and reliability.

Testing Focus and Activities

To establish data quality baselines, testing at the source phase focuses on validating data types, format consistency, referential integrity, and completeness. Profiling is also conducted to identify duplicates, null values, and outliers that may indicate data quality issues.

Key testing activities include:

- **Data Type and Format Validation:** Ensures fields such as dates, numeric values, and text strings conform to expected formats.
- **Mandatory Field Checks:** Confirms essential fields (e.g., email, phone number) are populated to prevent missing or incomplete records.
- **Referential Integrity Validation:** Ensures relationships between datasets remain intact, such as matching line-item structures in SAP purchase orders to master records.
- **Profiling for Duplicates and Nulls:** Identifies anomalies that could impact data usability in later stages.

Use Cases

CRM Systems

Ensuring that extracted customer records have valid email and phone number formats prevents downstream communication failures in marketing and sales automation.

Procurement

Validating SAP purchase orders by checking line-item structures against master records ensures compliance with procurement policies and prevents financial reporting discrepancies.

2. Data Ingestion Phase

The data ingestion phase marks the critical entry point for information to flow from operational and third-party systems into the ecosystem. At this stage, data is often diverse in format, frequency, and structure, ranging from structured relational databases to semi-structured APIs and unstructured logs. However, ingestion challenges like incompatible formats, corrupt or missing records, and schema violations can compromise downstream processes. Any errors introduced during this phase are amplified through the pipeline, potentially leading to processing failures, unreliable analytics, or disrupted decision-making. Rigorous testing during ingestion ensures data integrity, completeness, and conformance to schemas, laying a solid foundation for smooth transformations and analytics downstream. This proactive approach reduces rework, builds trust in the data, and minimizes delays.

Testing Focus and Activities

To ensure a smooth ingestion process, data testing focuses on verifying that raw data is accurate, complete, and compatible with downstream systems. Testing accounts for differences in data formats and schema expectations, addressing potential issues such as corrupt or duplicate records. Testing also ensures adherence to business rules and technical specifications for incoming data.

Key testing activities include:

- **File Format Validation:** Ensures files from sources like APIs (e.g., JSON, XML) or cloud storage platforms (e.g., Amazon S3, Azure Blob Storage) adhere to expected file types and naming conventions.
- **Field-Level Completeness:** Confirms all mandatory fields, such as key identifiers, are populated across structured data (e.g., MySQL, Oracle) and semi-structured formats.
- **Schema Matching:** Verifies that data from relational databases (e.g., SQL Server) or ERP systems (e.g., SAP) conforms to metadata specifications.
- **Null, Encoding, and Invalid Characters:** Prevents issues from data originating in unstructured formats (e.g., logs, multimedia files) or real-time data streams (e.g., Kafka, Kinesis).
- **Duplicate Detection:** Eliminates redundancies in batch file drops from cloud storage or duplicate entries in API pulls, ensuring data integrity.

Use Cases

Healthcare

Validating HL7 feeds ensures the structural integrity of patient demographics, lab orders, and discharge data. This is critical for real-time clinical decision-making and preventing adverse outcomes.

Insurance

Verifying broker-submitted claims (XML format) ensures schema compliance, reducing the risk of invalid claims disrupting claims processing and causing compliance breaches.

3. Data Staging and Transformation Phase

The data staging and transformation phase is a pivotal stage in the data pipeline, where raw data is processed into enriched, usable forms by applying business rules, performing lookups, cleansing, and calculating derived metrics. This step plays a crucial role in ensuring data meets business requirements and provides reliable analytics. However, this phase is also fraught with challenges. Missteps such as incorrect logic, faulty joins, and inconsistent business rule application can result in inaccurate aggregations, loss of data fidelity, and misrepresentation of metrics. These risks highlight the importance of rigorous testing, which ensures that transformed data accurately reflects business logic, maintains usability, and aligns calculated metrics with organizational objectives.

Testing Focus and Activities

Testing in the transformation phase emphasizes verifying that the data processing logic is technically sound and that business intent is preserved throughout the transformations. Since data transformations rely heavily on varied tools and engines such as ETL platforms (e.g., Informatica, Talend), ELT solutions (e.g., dbt, Snowflake SQL, Spark transformations in Databricks), and custom Python/SQL/Spark scripts, the testing approach must account for platform-specific execution semantics and transformation logic.

Key validations include:

- **Business Rule Validation:** Ensures consistent application of transformation logic (e.g., discount calculations or fraud detection) using tools like Informatica or dbt.
- **Lookup and Join Accuracy Checks:** Validates that dataset merging is accurate, and that foreign keys or enrichment values are correctly applied, leveraging platforms like Databricks Spark or custom SQL scripts.
- **Aggregation Validations:** Confirms that group-by summaries and totals are mathematically and contextually correct, with validations implemented in Snowflake SQL or equivalent ELT platforms.
- **Derived Field Validation:** Tests newly calculated fields, such as premiums, risk ratings, or interest rates, using Spark transformations or custom solutions.
- **Lineage Traceability:** Verifies that every output column traces back to its respective input fields and logic, ensuring transparency and accuracy.

Industry Use Cases

Banking

Testing ensures interest calculation logic adheres to account types and regulatory standards, preventing discrepancies and ensuring transparency during audits and customer communication.

Insurance

Validating premium adjustments based on customer risk profiles ensures pricing accuracy, maintains profitability, and avoids compliance breaches.

4. Data Storage and Warehouse Phase

The data storage and warehouse phase serve as the foundation for enterprise analytics and business intelligence. At this stage, data that has undergone transformation is expected to be clean, structured, and ready for query, but challenges such as duplicate records, incomplete dimensions, and unnoticed schema changes can threaten its reliability. These issues not only disrupt analytics but can cascade downstream, affecting dashboards, reporting, and AI models. Rigorous testing during this phase ensures structural integrity, historical accuracy, and consistency across dimensional relationships, making the data warehouse or lake a dependable source of truth for all business processes.

Testing Focus and Activities

Testing at this phase ensures that data stored in structured environments like data warehouses (e.g., Snowflake, Redshift, BigQuery) and flexible storage systems such as data lakes and lakehouses (e.g., Azure Data Lake, Databricks Delta Lake) is both technically and analytically sound. This includes confirming schema correctness, validating historical data completeness, and ensuring no redundancy or inconsistency in records.

Key testing activities:

- **Source-to-Target Reconciliation:** Ensures record counts, column totals, and random sample matches between source tables and loaded data.
- **Referential Integrity Validation:** Confirms relationships between fact and dimension tables are intact, verifying primary and foreign key dependencies.
- **Duplicate and Uniqueness Checks:** Detects and resolves duplicate entries to maintain integrity in primary and natural keys.
- **Partitioning Validation:** Validates completeness of incremental loads and ensures all expected organizational units or time windows are appropriately loaded.
- **Schema Drift Detection:** Monitors for unexpected changes in schema, such as dropped fields or altered data types, ensuring compatibility across downstream systems.

Industry Use Cases

Healthcare

Verifying that patient visit records map to unique identifiers ensures accurate clinical reporting, avoids duplicate entries, and prevents billing disputes. Testing safeguards against mismatched or redundant records, critical for regulatory compliance.

Banking

Matching transactional data with ledger balances post-load ensures the integrity of financial reporting and prevents discrepancies that could lead to internal audit findings or penalties.

5. Data Consumption and Reporting Phase

The data consumption and reporting phase is where data becomes visible to end-users and stakeholders through dashboards, reports, APIs, and advanced analytics models. This highly critical stage ensures that insights provided to users are accurate, timely, and consistent with underlying data. However, challenges such as misconfigured dashboards, delayed data refreshes, mismatched KPIs, or unvalidated model predictions can undermine trust in data and decision-making processes. In extreme cases, AI/ML pipelines consuming unreliable data can generate flawed forecasts or even unethical outcomes. By rigorously testing at this phase, organizations can uphold confidence in their metrics, empower data-driven decisions, and minimize risks from inconsistencies or biases in reporting and analytics.

Testing Focus and Activities

Testing during this phase focuses on ensuring that the metrics, aggregations, and data visualizations displayed to stakeholders accurately reflect the underlying data warehouse or lake. Special emphasis is placed on validating correctness, completeness, freshness, and secure access to sensitive data. Moreover, testing ensures the reliability of diverse consumption layers, such as dashboards (e.g., Power BI, Tableau, Looker), self-service analytics tools (e.g., Excel, Qlik), APIs, and AI/ML models.

Key testing activities:

- **Metric Consistency Checks:** Validates that computed KPIs and calculations in dashboards align with warehouse queries or business logic, ensuring decision-making reliability.
- **Timeliness and Freshness Verification:** Confirms data is up to date as per SLA or reporting frequency, ensuring insights are timely and actionable.
- **Visualization-Level Validations:** Ensures filters, slicers, and chart groupings operate correctly and reflect expected logic.
- **Cross-Platform Validation:** Confirms data consistency across multiple sources or BI layers, critical for composite reports.
- **Access Control Testing:** Verifies that users can only view or interact with data based on their roles and permissions, ensuring security and compliance.
- **AI/ML Model Validation:** Ensures that prediction pipelines consume verified data, preventing flawed outcomes in decision automation, risk scoring, or recommendations.

Industry Use Cases

Insurance

Dashboards displaying real-time policy lapse rates must accurately reflect backend data to allow timely interventions from agents or policyholders. Errors in metrics could lead to missed renewals, revenue loss, or reputational damage.

Banking

Loan disbursement metrics in daily reports must match transactional records in the warehouse, enabling trustworthy operational oversight and accurate forecasting. Inconsistencies at this stage could mislead executives and compromise business strategies.

Best Practices for Effective Data Testing

Implementing robust data testing requires a strategic, comprehensive approach. The following best practices provide a framework for organizations to transform their data quality management.

1. Understand the data landscape thoroughly.

A foundational best practice in data testing is developing a deep understanding of the data landscape before test design begins. This involves profiling both source and target systems to assess data formats, volumes, schemas, data relationships, and transformation logic.

In the insurance industry, for instance, inconsistent date formats and varying policy ID structures across internal and third-party systems can cause reconciliation errors in claims processing. By profiling the data early, the team can address these inconsistencies proactively, reducing downstream defects by considerable amount early in the data life cycle. Understanding the landscape ensures accurate test case design and minimizes the risk of failed data pipelines or inaccurate reports due to misaligned assumptions.

2. Define clear data quality and testing objectives.

Setting clear, measurable objectives is critical to guiding effective data testing. These objectives should focus on data quality dimensions such as completeness, accuracy, consistency, validity, and timeliness, and must align with business KPIs & requirements.

In a banking KYC remediation initiative, for instance, the data team should have an objective of 100% completeness of customer address information and zero duplication across customer records. These goals would help to ensure Anti Money Laundering (AML) compliance and improved trust in customer master data among frontline staff. Clear objectives not only provide direction but also help prioritize test scenarios based on business risk and compliance impact.

3. Develop a comprehensive data testing strategy.

Establish clear objectives and scope for structured, scalable data testing. The strategy should define responsibilities, tools, automation plans, validation techniques, and integration with the SDLC. It must align with enterprise architecture, business needs, and compliance standards while ensuring comprehensive test coverage and automation where feasible. A well-structured strategy enhances consistency, traceability, and predictability in data quality efforts.

For example, in a large-scale healthcare data modernization, a national provider migrating to the cloud must validate each data pipeline—from ingestion to transformation in ETL tools (e.g., Informatica) to reporting in Power BI. This includes layered testing, defining roles across QA, DataOps, and compliance, introducing data quality scoring, and mapping validations to business KPIs like claim turnaround time and patient eligibility accuracy.

A robust strategy prevents production incidents, ensures compliance, and reduces test cycles. It also serves as a blueprint for future initiatives, fostering collaboration and accelerating the path to high-quality, trusted data.

4. Establish a dedicated data testing team.

Form a team of qualified professionals responsible for ensuring data quality and compliance throughout the enterprise. This team should include data engineers, quality assurance specialists, and domain experts who can collaborate effectively with business stakeholders.

5. Leverage automated data testing tools.

Employ AI-driven frameworks to scale data validation and enhance efficiency. Automated tools can perform comprehensive testing across large datasets, identify patterns and anomalies that might escape manual detection, and generate detailed reports for continuous improvement. Automating data validations, schema checks, and transformation rule tests not only boosts efficiency but also enables continuous testing within CI/CD workflows for real-time assurance.

6. Ensure traceability.

Traceability and auditability are essential for validating test completeness and debugging issues. This includes linking test cases to business rules and maintaining detailed logs of queries, results, and test outcomes.

During data migration for a large insurance company, for instance, it is essential to trace every monetary value from the legacy system to the final financial report. This transparency helps satisfy internal audit requirements and boosts the confidence of finance teams and builds trust in the data testing process.

7. Use representative and secure test data.

Using test data that mirrors production patterns while ensuring privacy and compliance is critical in data-sensitive industries.

In healthcare, for example, synthetic data generation to simulate realistic patient demographics and treatment patterns without using actual patient records is essential to identify critical issues at the early stage of the data life cycle. Secure, representative data allows teams to detect issues that would surface in production, without violating regulations. It also reduces the risk of invalid test results due to incomplete or unrepresentative data sets.

8. Implement continuous data profiling and audits.

Regularly analyze and audit datasets to detect anomalies and inconsistencies. Good profiling of data by reviewing the source and understanding interrelationships in data is essential for maintaining your quality over time. Regular audits ensure ongoing compliance with internal standards and external regulations.

9. Adopt real-time data monitoring.

Continuous monitoring systems will enable proactive detection and resolution of your data issues. Real-time monitoring alerts stakeholders to potential problems before they impact business operations, analytics, or customer experience—allowing for immediate intervention.

Take a bank that has implemented data monitoring in data pipelines processing credit card transaction data, for example. When a vendor makes schema changes without notice, the system detects data loss and alerts engineers before inaccurate reports are published. Proactive monitoring reduces incident detection time significantly. Continuous monitoring ensures operational reliability, improves data trust, and minimizes the risk of data downtime in mission-critical systems.

10. Encourage cross-functional collaboration.

Foster synergy between data engineers, QA teams, business analysts, and compliance officers to strengthen data governance. Clear data stewardship is essential—assign data stewards to maintain and uplift data quality and initiate collaboration between your business and IT teams.

11. Integrate data testing into CI/CD pipelines.

Ensure high data quality throughout the lifecycle by embedding data validation in software development workflows. When data testing is integrated into development processes, potential issues are identified earlier, thus reducing the cost and impact of remediation.

12. Perform end-to-end data validation.

Verify data quality from ingestion to consumption to maintain consistency and accuracy across systems. End-to-end validation ensures that data remains accurate and reliable as it flows through different systems, transformations, and use cases within the organization.

13. Track metrics and continuously improve.

Tracking and analyzing metrics such as data quality rule failures, test coverage, defect density, and time to detect anomalies enables continuous improvement. These insights drive process improvements that cut defect leakage and shift the team from reactive to proactive issue management. By measuring what matters, organizations can evolve their data testing maturity and achieve long-term data quality excellence.

Structured data testing methodologies, automation, and continuous validation empower organizations to safeguard their data assets, drive business efficiency, and gain a competitive edge in the digital economy.

Achieve Data Integrity with Myridius

Data testing involves complex methodologies and strategic planning that require specialized knowledge and technical expertise to implement effectively. At Myridius, we offer end-to-end enterprise data integrity testing solutions powered by Tricentis that safeguard data reliability and business efficiency. Our experts design tailored strategies to address industry-specific challenges so your organization can establish compliance, accuracy, and trust in data-driven initiatives. When you partner with Myridius, your team can focus on core responsibilities while minimizing disruptions.

Here's how Myridius helps organizations maintain data integrity across key business functions:

Data Quality for Data Migration
When organizations migrate from legacy systems to modern cloud platforms, data testing ensures the accuracy and completeness of migrated data.

Data Quality for Application Modernization
During application modernization and platform transitions, data testing validates that all data maintains its integrity throughout the transformation.

Data Integrity for Compliance
Regulatory requirements such as GDPR and CCPA require organizations to manage and be accountable for the data they collect and process. Data testing ensures compliance by validating that all data meets regulatory standards and privacy requirements, reducing the risk of penalties and enhancing trust in organizational data.

Data Integrity for AI Readiness
AI models and their results are only as reliable as the data they're trained on. Our proven approach to data testing prepares organizations for AI adoption by ensuring the quality, completeness, and appropriateness of training data.

No matter the use case, a proactive approach to data testing enables organizations to mitigate risk, improve efficiency, and leverage their data assets. With Myridius, your data remains a trusted foundation for innovation, compliance, and strategic decision-making.

Build a Future-Ready Data Strategy

In an AI-enabled, digital-first world, data quality is essential for long-term success. Myridius provides end-to-end data integrity solutions to help businesses tap into the full potential of their data. Don't let poor data quality put your business at risk—safeguard your success with Myridius today.

BOOK A MEETING

myridius.com | (732) 744-3500 | info@myridius.com