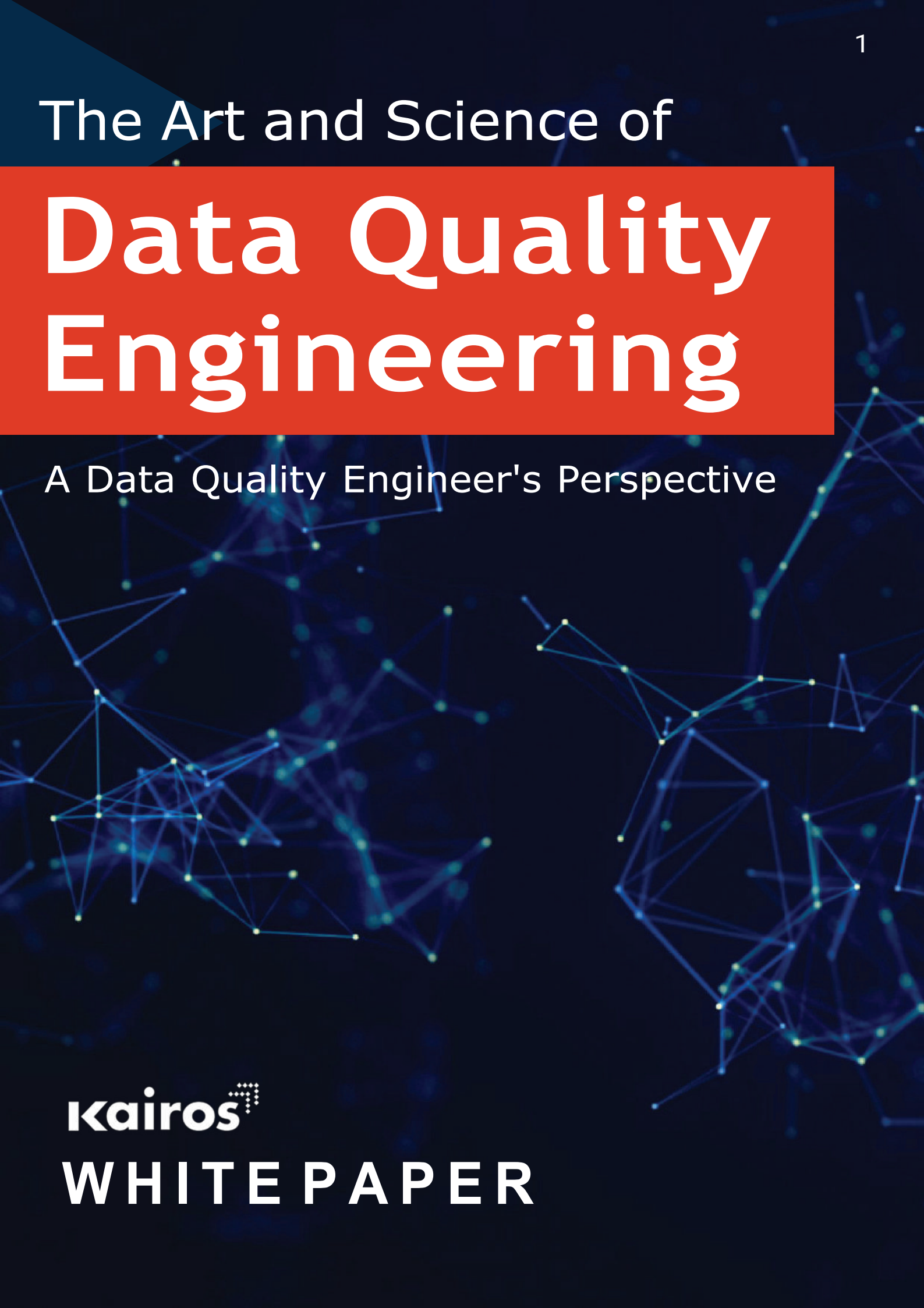
The Art and Science of Data Quality Engineering

A Data Quality Engineer's Perspective



Kairos

WHITE PAPER

Kairos

Enabling Extraordinary Experiences

ABSTRACT

This paper will explore the solutions, opportunities, and technological innovations that ensure data quality across business verticals.

Data quality becomes increasingly important as the world moves towards a more digitalized world. If data is incorrect or inaccurate, it can have severe consequences for businesses and individuals. Simply put, data quality is the quality of information. It refers to how accurate, complete, and reliable a piece of information is. Poor data quality can refer to simple errors (typos) or more complex issues (like inaccurate business data).

This paper discusses data quality concisely, outlining the various concepts and techniques used to improve data quality. It also provides a comprehensive overview of modern data quality engineering, highlighting the different applications and methods that are currently in use. This paper then looks at how Kairos Technologies solutions can help enterprises looking to meet or exceed data quality needs and establish a core platform that adds flexibility to meet future change and growth. By understanding data quality engineering, you'll be better equipped to make informed decisions regarding data management.

What's inside

IMPORTANCE OF DATA QUALITY IN THE DIGITAL WORLD

The significance of data quality cannot be oversimplified. With more data comes a flood of new information and data heterogeneity. The massive amount of data must be managed to make any sense to drive informed decision-making and take calculated risks. The digital world is becoming increasingly reliant on data to function. Data can range from user behavior data to product data.

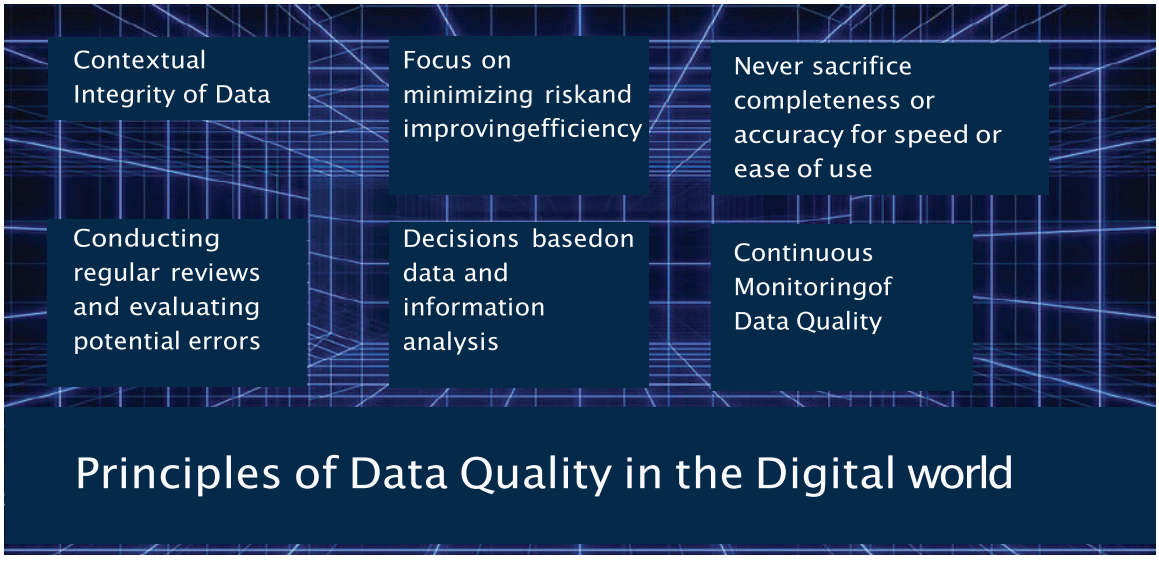
Accuracy is one of the important dimensions of data quality for data-driven businesses in the modern digital world. Data Quality must be maintained to ensure reliable and accurate information.

Incorrect expectations, missed opportunities, incorrect search engine results, inaccuracies in business decisions, vulnerability to cyberattacks, and wasted time and money can all be impacted by poor data collection, processing, or storage.

Poor data quality can harm a company's reputation, cause customer loss, and reduce employee productivity.

Organizations must take several steps to maintain data quality, including ensuring that data is accurate and up to date and using appropriate software for data capture, storage, transfer, and consumption.

Maintaining high-quality data is a constant challenge, but it can be tackled with the right tools and processes. This is especially important in today’s world, where so much happens through digital channels.



DATA QUALITY ENGINEERING
a Comprehensive Guide

Data Quality Engineering (DQE) ensures that data are accurate, reliable, and usable. It includes activities such as inspecting data for consistency and completeness; correcting data where necessary; verifying the accuracy and performing testing to ensure proper functionality. Data Quality Engineering principles work together to produce high-quality datasets that meet business needs. Here are the essential components of DQE and how they work together to improve data quality:

Quality Assurance (QA)

Ensuring all aspects of the data collection process are followed to minimize errors or inconsistencies in the dataset. QA includes reviewing and correcting data where necessary and verifying accuracy.

Data validation

Ensuring the data entered a system is accurate and error-free. Data validation includes checking for inconsistencies, incorrect or missing information, and out-of-date information.

Quality control (QC)

Checking the dataset's quality throughout its life cycle to ensure that it meets predetermined standards. QC includes inspections and tests during development, testing after it has been integrated into a system, and monitoring and managing datasets to ensure they meet requirements.

DQE is an integral part of ensuring the quality of data.

The following principles guide DQE:

Data quality first — Ensuring that all data collection, use, and disposition decisions are based on achieving the desired business objectives.

A holistic approach to data management — A systems-based approach to managing data ensures that all aspects of the data collection process are considered.

Data accuracy first — Taking measures to ensure that data entering a system areaccurate and error-free from the outset.

Open and transparent data management — Making data available in a format that is useful and accessible to both internal and external stakeholders.

Continual improvement of data quality —Ensuring that data quality is continuouslyimproved through rigorous testing, evaluation, and adjustment.

Responsible use of data —Taking measures to ensure that it is used responsibly and consistent with the objectives for which it was collected or generated.

Data governance —Ensuring that all appropriate policies and processes are in place to protect datasets and enforce data quality and accuracy standards.

A Data Quality Assurance Platform **(DQAP)** helps organizations manage their data quality process and track the effectiveness of their DQE strategies. There are various features that a DQAP can have, including:

Data validation tools allow users to approve or reject data entries as they enter them into the system to ensure that they meet specific requirements.

Data quality assessment tools provide an overview of the overall quality of a dataset based on defined criteria. They can also help identify potential issues early on before they become too difficult or expensive to fix.

Data management tools allow users to manage and use data in various ways, including exporting it into different formats, tracking changes over time, and linking it to other data sources.

Data cleansing and quality control tools help remove inaccurate or erroneous data and unwanted material.

Reporting and analysis tools allow users to explore the data in various ways, including drilling down into specific areas, looking at trends over time, or creating charts and graphs.

Data governance tools help to manage who has access to the data, how it is used, and under what conditions.

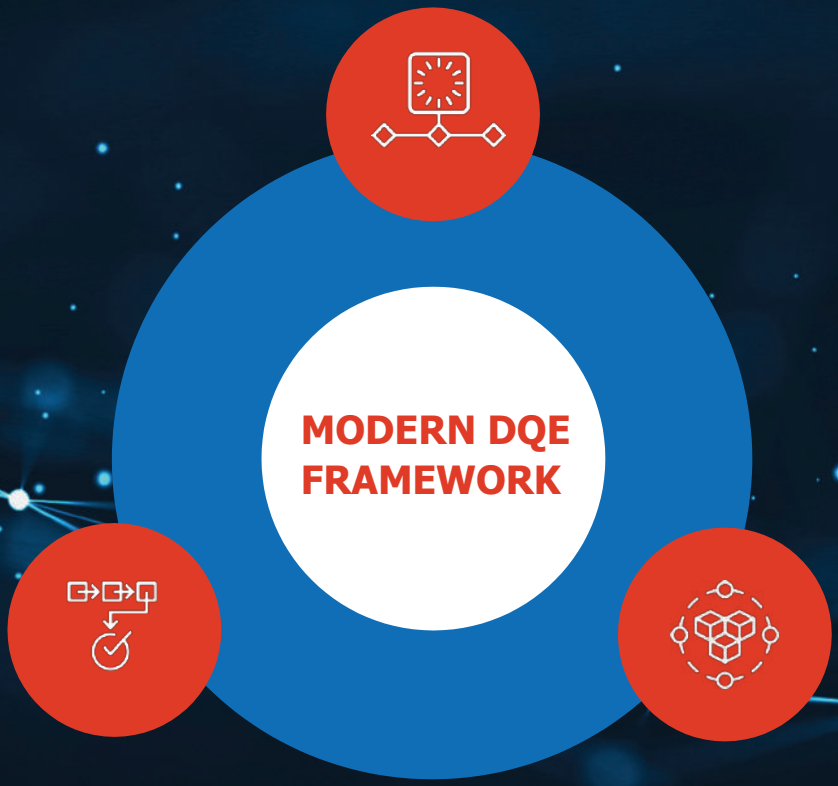
Data quality assurance support offers advice, assistance, and resources to help ensure the accuracy and integrity of data.

"Additionally, various analytical tools are used to detect, understand, and rectify flaws in data to allow for efficient data and analytics governance throughout operational business processes and decision-making. The available tools and solutions address a variety of essential responsibilities, such as profiling, parsing, standardization, cleansing, matching, monitoring, rule creation and analytics, as well as built-in workflow, knowledge bases and collaboration."

The Modern Data Quality Engineering Framework

PHASES

- Pre-Processing
- Data Integration and Validation
- Quality Assurance
- Reporting



Approach

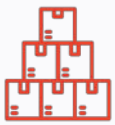
- Identify Issues
- Comprehensive Solutions
- DQ & QA Mechanisms
- Measuring the effectiveness of data quality initiatives

MODULES

- Data Quality Assessment
- Data Acquisition
- Data Governance
- Data Disposition
- Process Improvement
- Data Preparation

The Key Benefits of DataQuality Engineering

Make Better Decisions



Improve Data Quality

Covers Data Quality life cycle

Increase Competitiveness

Increase Performance

Increase Data Usability

Increase Customer Satisfaction

DQE empowers a seamless end-to-end journey to the desired outcomes

THE DIFFERENT TYPES OF DATA QUALITY ISSUES AND HOW TO FIX

Inaccurate Data: This may include incorrect information about customers, orders, products, or any other aspect of your business. To correct this issue, you must gather accurate data from all sources and ensure that it is consistently entered into your systems. You can also use data cleansing techniques to remove inaccurate information from your database.

Data That Is Out of Date or No Longer Applicable: This may occur when old files have been deleted, or entries have been changed since they were last used; these changes might not be reflected in current versions of databases and applications. You must update data access mechanisms and software applications to correct this problem.

Excessive Processing on Data: This occurs when extreme calculations are performed on data before they're used in analysis or decisions, which can distort the information presented. To avoid this, use appropriate algorithms and processing methods when accessing and using data for analysis.

Lack of Business Accountability: If data is generated without any agenda, it could lead to incorrect conclusions about the information in a system. To prevent thisfrom happening, establish DQ responsibilities and operating procedures and engage Data Stewards and Data Custodians across BUs and IT.

Inability to Repeat Results: When data is constantly changing and not repeatable, it can be challenging to find reliable conclusions or trends. To avoid this, ensure all data is regularly reproduced and evaluated so that decisions can be drawn with consistency and reliability.

Duplicate Data: Errors may occur when duplicate data is mistakenly created or entered a database. This can lead to inconsistencies and confusion within the system and potential issues with system performance. To correct this problem, you must reconcile the data based on appropriate de-duplication rules.

Incorrect Formatting: Data that is improperly formatted can cause difficulties in processing or displaying it on screens. This may include inaccurate information, duplicate data entries, and typos. To correct this issue, you will need to reformat the data using standard database formatting rules and procedures.

Inconsistencies: Include variations in values that don't seem consistent with each other. You can fix this issue by reconciling different data sets. The first step in reconciling different data sets is to identify the similarities between the data sets. Once the similarities are identified, the data sets can be reconciled by using the similarities to fill in the missing data.

Logical Errors: Incorrect relationships between pieces of information are part of a Logical error. The correlation does not imply causation. Correlation is a measure of the strength of the relationship between two variables. It does not mean that one variable causes the other. Such errors can be corrected by testing assumptions made about the data.

Kairos makes data quality management easy and seamless

“The competitive edge of any organization comes down to its data quality. As technologies evolve and data quality issues continue to grow, companies need to understand what customers want, how well products are performing, and the right steps to take to respond effectively to the demands of the rapidly changing marketplace.”

Kairos' next-generation, cloud-based, and on-premise, "DQGateway," is a no-code, AI/ML-powered visual Data Quality platform. The platform seamlessly connects to multiple data sources, decreasing the complexity of data ecosystems providing reliable decision-making, and making your data a profitable asset. With DQGateway, organizations can be confident that they have accurate, timely data to make informed decisions and stay ahead of the competition.

Our flagship DQGateway, a data quality tool, is designed specifically for general-purpose data quality applications. It delivers core data quality functions such as profiling, interactive visualization, business rule creation, rule-based data validation, parsing, standardization, cleansing, matching, and multidomain data support. With DQGateway, you can quickly evaluate, monitor, and manage the quality of data in your information systems. And because it comes loaded with specific sets of rules, Kairos DQGateway is the perfect solution for public and private sector organizations that manage large volumes of data.

So, if you're looking for an easy-to-implement data quality management solution that can provide tangible business gains, look no further than Kairos Technologies.

Features of DQGateway

- No-code solution for automation of ETL testing and the gateway to achieving Continuous Data Quality Monitoring.
- Data profiling with automatic detection and masking of sensitive data fields.
- DataOps with Continuous Testing (Jenkins integration).
- Data Quality Dashboard. Email notification of test results based on subscription.
- An easy-to-use interface with plug-and-play connectors that integrate into the DQGateway ecosystem allows rapid data quality deployment.
- Define rules and automated processes for data quality improvement.
- Visual Check Mate for automated BI report validation.
- Unattended data quality checks using scheduled execution.
- Reports (with PDF download option).
- A set of algorithms adept at hierarchical unification by identifier keys irrespective of internal data structures - can perform approximate matching in record unification.
- Configurable Data Quality checks for - Completeness, Correctness, and Comparative Reconciliation.

In addition, our tool offers a variety of data quality rule engines, plug-ins, and integrations that can be tailored to your specific needs. And because we understand the needs of our clients, our solutions are architected to be extensible to meet those specific needs. Please visit our website or contact us today to learn more about Kairos Data Quality Gateway.

DQGateway

Making Data quality management easy

CLOUD

DQGateway is the platform that supports your Cloud storage.

Built-in data connectors that can access data stored in any file format, database, cloud store, or obtained from application APIs. All controls are from one location. Using out-of-the-box connectors, attach to a wide variety of data sources such as on-premises or cloud databases, data warehouse, files, and storage such as MySQL, SQL Server, Snowflake, Excel, CSV, JSON, XML, and more.

OPEN

Get the total value from the open-standard solution

The solution is easily configured with administration applications. You will not need any external tools or 3rd party applications. Extensible and customizable for specific client requirements. Discover meaningful insights from the data profile views. These come with automatic masking of sensitive data.

Quality

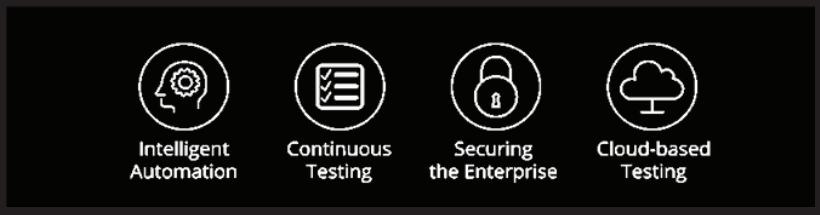
A priority at every step

Completeness, Correctness, Comparative Reconciliation. Measure Data Quality - Configurable set of built-in quality checks can be combined into powerful Data Quality Gatekeepers. These can be invoked interactively from pipelines or run unattended to provide a score for different data sources—a client-specific data quality dashboard.

FLEXIBILITY

Adoption at a pace that is right for you

A cloud services model with On-Premises and Cloud deployment options backed by continuous enhancements and support.



The Future Building Blocks

About Kairos Technologies

Kairos technologies develop tools that simplify complex processes and create value in any industry. We build software to help engineers to develop products faster and better while reducing their risk of failure. You can make anything and build better. We aim to solve problems before they make it into your production environment.

Kairos provides quality engineers a platform to share knowledge, collaborate on projects, and track their work in real-time. We deliver simplified quality engineering services, helping our customers improve their products through better code coverage, faster development cycles, and higher performance. Our technology automates repetitive manual tasks, enhances efficiency and productivity, and prevents errors from being introduced into your live production environment. Learn more at www.kairostech.com.

Contact:

Radhika Rao
Chief Delivery Officer
Kairos Technologies

Team Kairos

Parthasarathi Bhattacharjee
Head of Innovation | Kairos Technologies
Parthasarathi.b@kairostech.com

Jaya Krishna Manchala
Sr Technical Content Writer | Kairos Technologies
Jayakrishna.m@kairostech.com