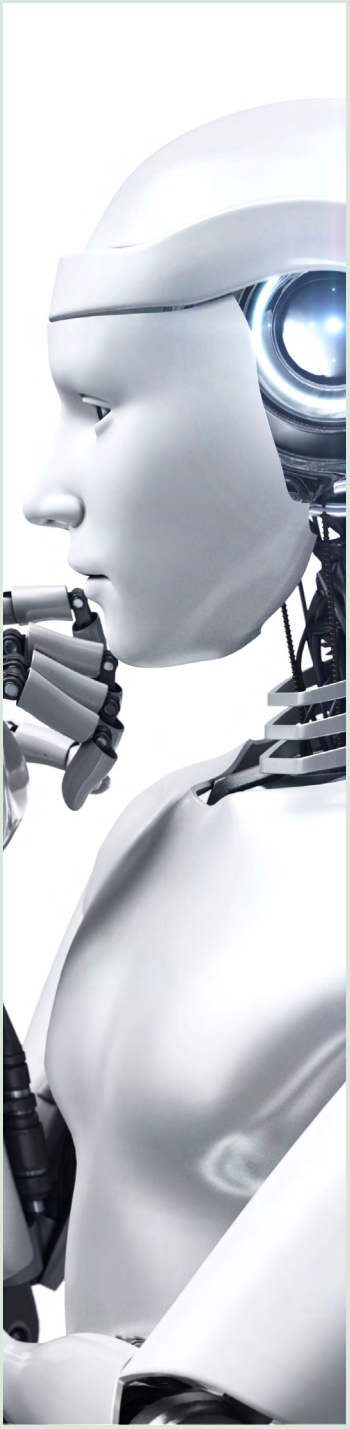
Infosys topaz

WHITE PAPER



LEVERAGE THE POWER OF AI FOR BETTER DATA QUALITY IN TRANSFORMATIONAL PROGRAMS

Infosys®
Navigate your next

As digital transformation and cloud computing become integral to post-pandemic global markets, businesses face the critical challenge of ensuring data quality for massive data migrating to the cloud. Data scattered across diverse systems requires enterprises to spend considerable time and effort stitching data before reporting and analysis. This requires an AI-first data quality (DQ) framework to address the issue. This paper explores the relevance and potential of an AI-driven data quality framework to revolutionize data testing practices. It describes the technology involved in the three phases of the cloud data journey. It also highlights the advantages of an AI-powered DQ framework, including enhanced testing coverage, cost savings, and faster time-to-market. By complementing existing automation frameworks, AI opens newer frontiers to data-led quality checks and more effective data transformation projects.

Introduction

Organizations today are actively exploring various ways to integrate AI-based testing practices with traditional manual data assurance approaches, leading to the emergence of a hybrid assurance model. The progression varies in maturity, from adopting simple machine learning (ML) models for early sanity tests to leveraging advanced deep learning neural networks and large language models (LLMs).

At the forefront of this transformative journey lies the development of an AI-first DQ framework, marking a crucial milestone in establishing a comprehensive AI-based data testing approach. Leveraging the latest advancements in generative AI, this framework has the potential to revolutionize test planning activities by replacing complex SQL script generation with natural language capabilities.

Neural network algorithms can analyze data to proficiently extract relationships and business rules from enormous volumes of data, while deep learning algorithms analyze extensive datasets to identify data patterns and detect potential defects misaligned with the data patterns. This data-driven evolution is propelling the realization of a fully automated AI-first DQ framework equipped with self-healing features that proactively identify data issues based on past learning and apply appropriate fixes before the issues escalate.

Relevance and Need

The traditional data testing approach relies heavily on manual effort, requiring significant human intervention during test script preparation. The tester's domain expertise and technical knowledge play a vital role in determining the quality of the test scripts. However, the conventional approach faces several roadblocks, including challenges in scaling for cloud transformation, extended time-to-market, and limited test coverage through manual means.

Addressing these critical challenges and ensuring enhanced data quality, the AI-first DQ framework emerges as a transformative solution for the cloud data journey. It ensures superior data quality by leveraging the power of AI to offer more efficient and effective methods of identifying and resolving data quality issues. It empowers data validation with qualitative test data generation while also automating monitoring and operations of data pipelines, leading to streamlined processes.

Gartner® expects: "By 2025, 30% of outbound marketing messages from large organizations will be synthetically generated, up from less than 2% in 2022." The prediction aligns with the concept of generating synthetic data in the AI-first DQ framework, indicating the growing importance of AI-driven data practices in various domains, including marketing.

Technology and Solution

The AI-first DQ framework is designed as a comprehensive AI-based data testing approach, catering to the distinct requirements of the three phases of the cloud data journey.

These include:

1. Data testing lifecycle
2. Test data management (TDM) and data masking
3. Data quality monitoring and operations

The following sections explore how each phase benefits from the transformative capabilities of the AI-first DQ framework.

Phase I: Data Testing Life Cycle

The conventional data testing lifecycles entails, test planning phase to plan for data migration, testing of migrated data and business rules associated with the data. Data represented in business dashboards are tested and finally overall data quality is certified for the entire data migration program.

Figure 1 provides a high-level overview of how an AI-driven test approach impacts various stages in the first phase of data migration testing.

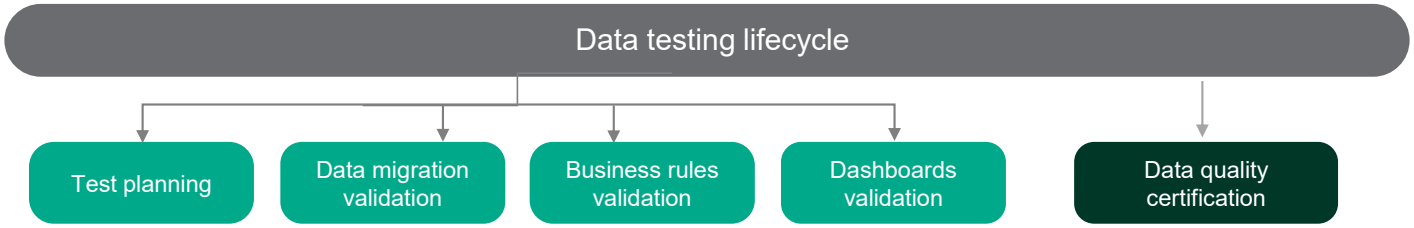
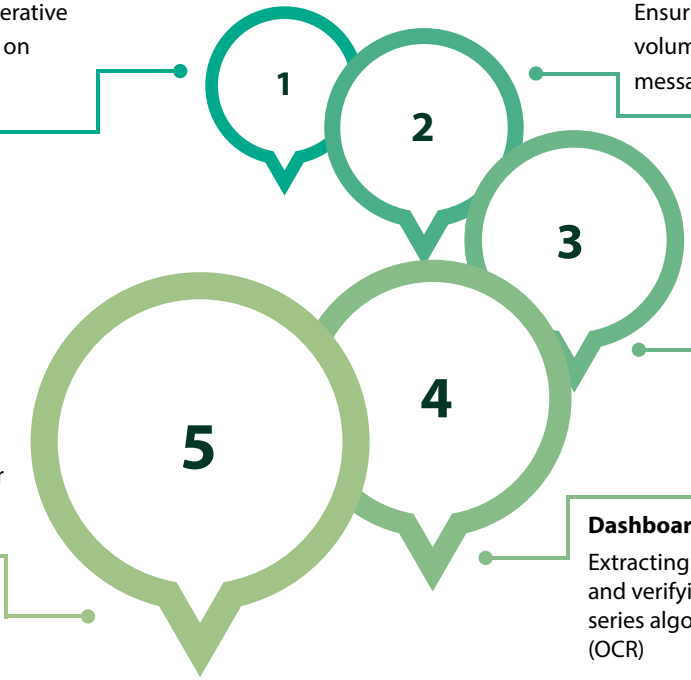


Fig 1: Overview of the data testing lifecycle

The data testing lifecycle encompasses a range of critical activities that play a pivotal role in ensuring data quality and reliability at each stage of data management. The many ways in which a data testing lifecycle can benefit from an AI-powered test approach are listed below:

Test planning
Leveraging the capabilities of generative AI to create SQL test scripts based on requirements

Data migration assurance
Ensuring quick sanity assurance for large volumes and varieties of data using hash/ message-digest5 (MD5) algorithms



Real-time DQ testing
Proactively identifying anomalies in data schemes and data profiles to prevent data failures through the application of anomaly/outlier detection algorithms

Data transformation testing
Extracting business rules from data and identifying records that do not meet the rules by using pattern analysis/deep learning algorithms

Dashboard testing
Extracting data from visualization dashboards and verifying it against data sources using time series algorithms/optical character recognition (OCR)

Phase II: Test Data Management and Data Masking

AI has evolved to address the challenges in the various stages of TDM and data masking to unlock a highly efficient and effective data transformation process. Figure 2 represents the TDM and data masking stage.

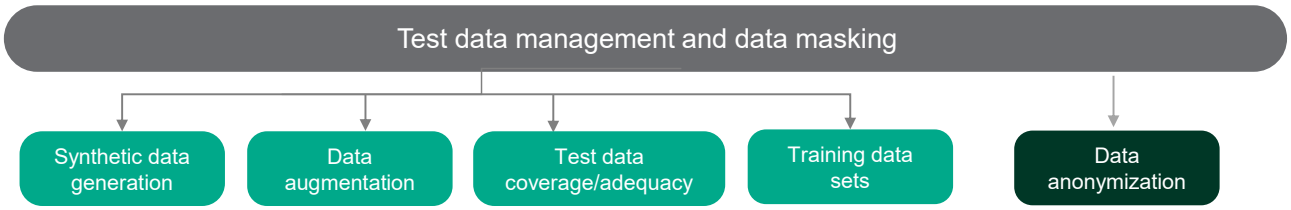


Fig 2: Overview of the test data management and data masking processes

By harnessing the advanced capabilities of AI, businesses can streamline data management practices, enhance data security, and optimize their data transformation journeys.

Here are some of the ways AI can help in TDM and data masking

1

Synthetic data generation

Leveraging long short-term memory models (LSTM) to generate production-like test data by learning patterns

2

Data augmentation

Using the Gaussian mixture models to augment or enrich existing test data

3

Test data coverage/adequacy

Employing greedy/genetic algorithms to analyze production data patterns against test beds, identifying gaps in test coverage, and optimizing test beds

4

Training datasets

Generating high-quality training datasets for model testing using the Synthetic Data Vault (SDV)

5

Data anonymization

Securing sensitive production data in test environments using artificial neural networks (ANN) for anonymizing or masking

Phase III: Data Quality Monitoring and Operations

AI has revolutionized the data quality monitoring process throughout the data journey. Figure 3 illustrates the data quality monitoring and operations phase in the data journey.

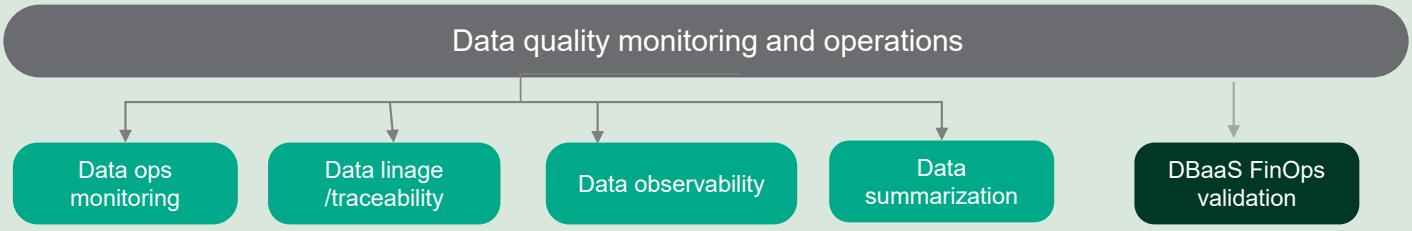


Fig 3: Overview of the data quality monitoring and operations stage

By utilizing the appropriate algorithms and models at each stage, businesses can automate the data quality monitoring process effectively in the following ways:

DataOps monitoring

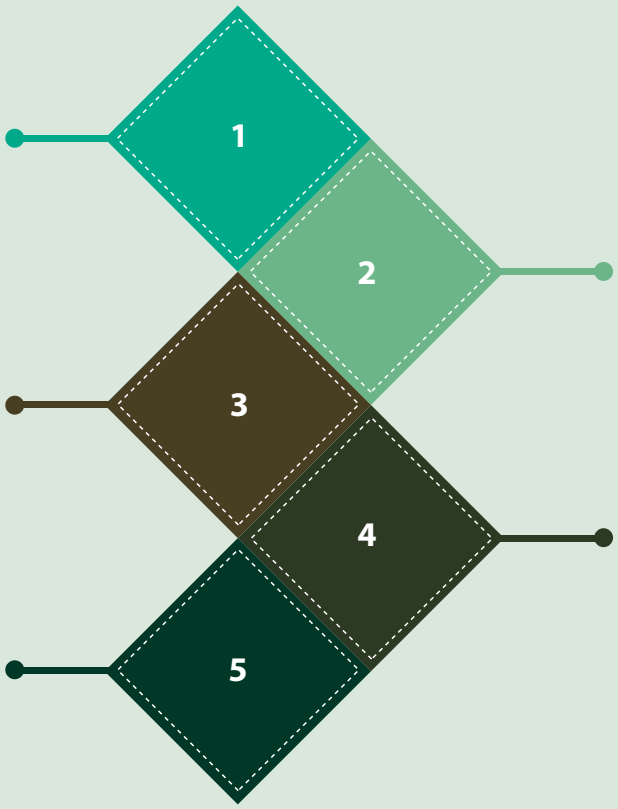
Monitoring of data pipelines to detect early warning signals or missing service level agreements (SLAs) and identifying data issues proactively

Data observability

Employing a self-learning prediction model to proactively identify data quality and data pipeline issues while leveraging it for self-healing capabilities

Database-as-a-Service (DBaaS) FinOps testing

Utilizing AI models to analyze cost usage reports from cloud hyperscalers, recommending configuration optimizations for storage cost savings



Data lineage/traceability

Tracing the source and lineage of data causing issues as well as using visual representations to identify the processes impacted by the incident

Data summarization

Leveraging generative AI summarization models to derive meaningful insights from the semantic
layer data

AI-first DQ framework developments are currently in their initial stages in the market. Although there are no existing market tools, solutions, or frameworks to leverage AI for comprehensive data checks, the industry is progressing well in adopting emerging technologies for data quality assessments. By integrating AI with existing automation frameworks, data-led quality checks have delivered remarkable advantages, including:

- Eliminating human errors
- Faster time to market
- Reducing test timelines significantly
- Increasing cost savings
- Ensuring the availability of quality test data for early detection of defects
- Providing better test coverage
- Enhancing compliance and avoiding data breaches
- Empowering self-service and self-healing capabilities
- Implementing a fail-fast fail-safe approach toward cloud data migration

Conclusion

The well-defined development process for AI in enterprises marks a crucial step, and yet the true potential lies in operationalizing AI within the software development lifecycle, leading to a paradigm shift towards continuous delivery of AI-based systems. Amidst the heightened activity and excitement surrounding AI in software testing, businesses are at the cusp **of a new era within the realm of test automation.** AI testing has empowered testers, developers, and software professionals to overcome challenges once considered insurmountable. The rapid advancements in the evolution of LLMs have accelerated the path towards a comprehensive A**I-first data quality framework, fostering efficient quality engineering practices for data transformation programs.**

References

Gartner, "Beyond ChatGPT: The Future of Generative AI for Enterprises", Contributor: Jackie Wiles, January 26, 2023.

Author

Saju Joseph

Digital Solution Specialist

Saju Joseph brings nearly twenty-three years of experience in different techno- functional roles. He has played the role of solution specialist for multiple digital transformations programs for clients worldwide, across industry verticals. He is currently based out of India and is responsible for developing new service offerings/platform solutions for AI, Data and Analytics Quality Assurance across industries. He holds a bachelor's degree in mechanical engineering from the prestigious Calicut University.Outside of his professional life, he enjoys travelling, reading, and spending time with his family.

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises and communities to create value. With 12,000+ AI use cases, 150+ pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems.

For more information, contact askus@infosys.com

Infosys®
Navigate your next