SOLVING THE TEST DATA CHALLENGE TO ACCELERATE DIGITAL TRANSFORMATION

Abstract

Organizations are increasingly adapting to the need to deliver products and services faster while continuously responding to market changes.

In the age of mobile apps, test automation is not new. But traditional test data management (TDM) approaches are unable to help app development teams address modern delivery challenges. Companies are increasingly struggling to keep up with the pace of development, maintain quality of delivery, and minimize the risk of a data breach.

This white paper illustrates the need for a smart, next-gen TDM solution to accelerate digital transformation by applying best practices in TDM, zero-trust architecture, and best-in-class test data generation capabilities.

Infosys®

Navigate your next

Table of Contents

Table of Figures

Traditional Test Data Management

Test data management (TDM) should ensure that test data is of the highest possible quality and available to users. In the digital age, managing test data using traditional TDM practices is challenging due to its inability to accelerate cloud adoption, protect customer data, provide reliable data, avoid data graveyards, ensure data consistency, and automate and provision test data.

Why the New Normal was not Enough?

While the 'new normal' has become a catchword in 2021, in the world of testing, this 'normal' was not effective for many organizations. The pressure to adapt to changing customer expectations, new technology trends, changing regulatory norms, increased cybersecurity threats, and scarcity of niche skills has raised many challenges for organizations. In light of this, many are wondering whether they should revisit their test data strategy.
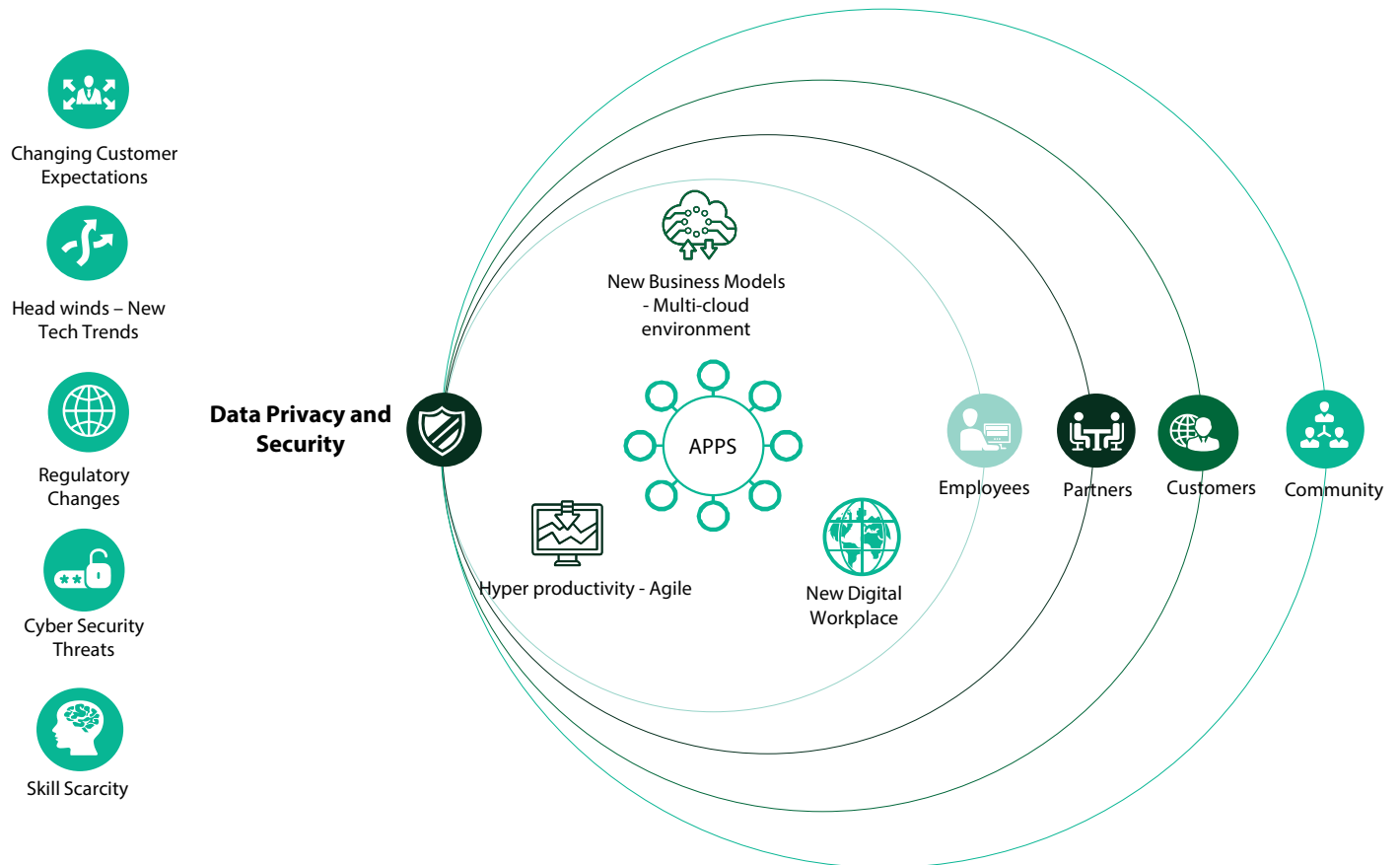

Figure 1. Key focus areas emerging in test data management

As time-to-market for products and services becomes critical, test data generation and provisioning emerge as bottlenecks to efficiency. Further, test data management has been represented as the weak link for organizations looking to accelerate digital transformation through continuous integration and delivery. High quality test data is a prerequisite to train machine learning (ML) models for accurate business insights and outcome predictions.

To build a competitive difference, organizations today are investing in three key focus areas in test data management (refer Figure 1):

- **New business models** – With a strong focus on customer experience, organizations must adopt new business models and accelerate innovation. There is a need to generate data that can be controlled and is realistic as well as accurate to meet real-world production needs.

- **Hyper-productivity** – Automation and iterative agile processes push the need for better testing experiences with faster and more efficient data provisioning, allowing organizations to do more with less.

- **New digital workplace** – Millions of employees are working from home. Organizations must focus on building a secure, new-age digital workplace to support remote working.

Five Key Drivers and Best Practices in Test Data Management

Companies are increasingly struggling to keep up with the pace of development, maintain quality of delivery, and achieve absolute data privacy. On-demand synthetic test data is a clear alternative to the traditional approach of sub-setting, masking, and reserving production data for key business analytics and testing. In this context, three key questions to ask are:

1. What are the drivers and best practices to be considered while building a test data strategy?

2. How can CIOs decide what is the right direction for their test data strategy?

3. What are the trade-offs in test data management?

There are five elements – cost, quality, security and privacy, tester experience, and data for AI – that drive a successful test data management strategy. Understanding the best-practices around these will guide CIOs in making the right decision.
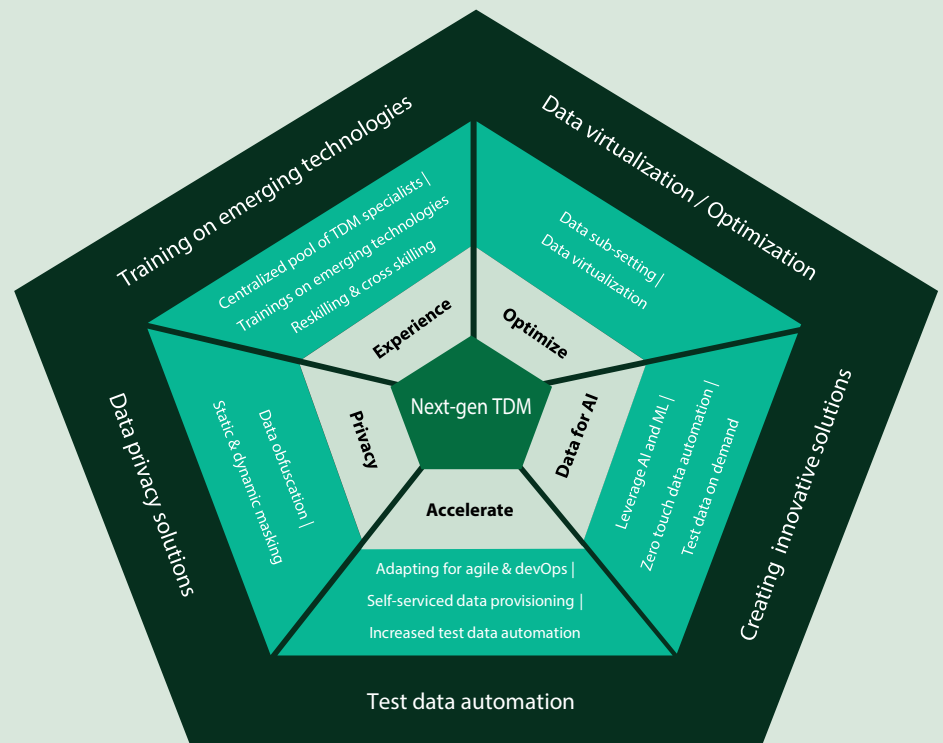

Figure 2. Key drivers and best practices in TDM

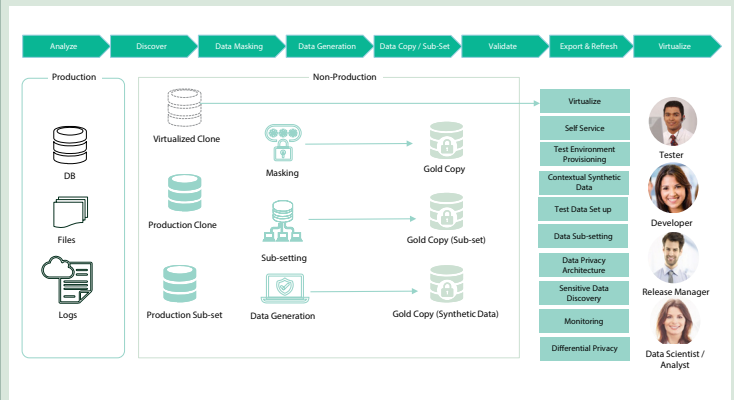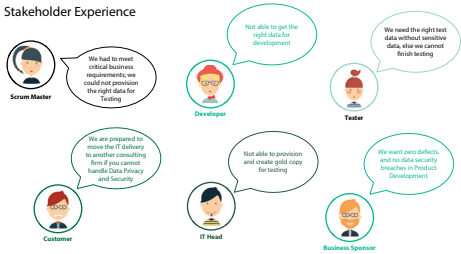| Key Drivers	Impact on Test Data Strategy	Best Practices
1. Cost	What is the return on investment (ROI) and acceptable investment to create, manage, process, and, most importantly, dispose of test data? Production data must be collected, processed, retained, and disposed of. The processing and storage cost must offset the investment in TDM products. Procurement, customization, and support costs need to be considered.	• Test data as a service – Test data on cloud with a subscription for testers can lower the provisioning of full-scale TDM.
2. Quality	Do we have the right quality of data? Can we get complete control over the data? Can we generate test data in any format? Testers have very limited control over the data provided by production. The test data is usually a subset of data from production and cannot cater to all the use cases including negative and other edge use cases. Further, there is a need to generate electronic data interchange (EDI) files, images, and even audio files for some of the use cases.	• TDM suite can help build a subset of data designed with realistic and referentially intact test data from across the distributed data sources with minimal cost and administrative effort. • Synthetic data generators should have the breadth to cover key data types and file formats along with the ability to generate high-quality data sets, whether structured or unstructured, across images, audio files, and file formats.
3. Security and privacy	Do we have the right data privacy controls while accessing data for testing? How do we handle a data privacy breach? The focus on privacy and security of the data used for testing is increasing. Complying with GDPR and ensuring the right data privacy controls is a catalyst for organizations to move away from using direct production data for testing purposes. There is increased adoption of masking, sub-setting, and synthetic data generation to avoid critical data breaches when using sensitive customer, partner, or employee data.	• Zero trust architecture provides a data-first approach, which is secure by design for each workload and identity-aware for every persona in the test management process including testers, developers, release managers, and data analysts. Figure 3. Zero trust architecture • To ensure security of sensitive information, organizations can create realistic data in non-production environments without exposing sensitive data to unauthorized users. Enterprises can leverage contextual data masking techniques to anonymize key data elements across an enterprise.

4. Tester experience	Are we building the right experience for the tester? Is it easy for testers to get the data they need for their tests? Customers struggle to meet the agile development and testing demands of iterative cycles. Testers are often forced to manually modify the production data into usable values for their tests. Teams struggle to effectively standardize and sub-set the production data that has been masked and moved to test data. Figure 4. Stakeholder experience	• Test data automation puts the focus on tester experience by enabling a streamlined and consistent process with automated workflows of self-service test data provisioning • Test data virtualization allows applications to automatically deliver virtual copies of production data for non-production use cases. It also reduces the storage space required.
5. Data for AI	Do we understand insights generated by the data? The probabilistic nature of AI makes it very complex to generate test data for training AI models.	• Adopt mechanisms for data discovery, exploration and due diligence. Data resides in different formats across systems. Enterprises must identify patterns across multiple systems and file formats and provide a correct depiction of the data types, locations, and compliance rules according to industry-specific regulations. They should also focus on identifying patterns, defects, sub-optimal performance, and underlying risks in the data. • For data augmentation, analysts and data scientists can be provided with datasets for analysis. The datasets must be resistant to reconstruction through differential privacy for effective data privacy protection.

Future-proof Test Data through Next-gen TDM Innovation

Every organization needs simplified testing models that can support a diverse set of data types. This has never been a higher priority.

Infosys Next-Gen TDM supports digital transformation by focusing on 9 key areas of innovation (see Figure 5). The offering leverages the latest advances from data science in test data management, giving enterprises the right tools to engineer appropriate test data.
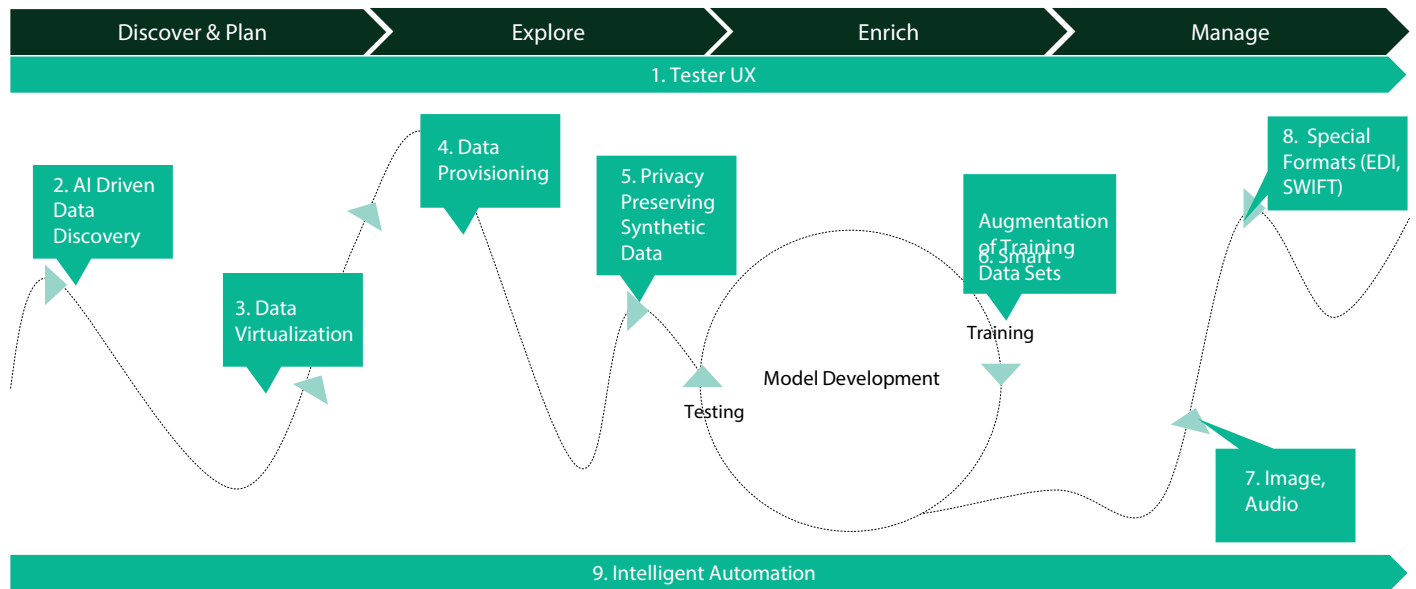

Figure 5. Focus areas of Infosys Next-Gen TDM

1. **Tester user experience** – Testers need to assess business and technical requirements from the perspective of testability as well as end users. Infosys Next-Gen TDM provides a framework that includes testers and gives them a 360-degree view of the TDM process.

2. **AI-driven data discovery** – Modern test data resides on a tower of abstractions, patterns, test data sources, and privacy dependencies. One of the key features of Infosys Next-Gen TDM is smart data discovery of structured and unstructured data using AI. This helps uncover:

   • Sensitive data (PII/PHI/SPI) to avoid data privacy breaches

   • Data lineages to build the right contextual data while maintaining referential integrity across child and parent tables

3. **Data virtualization** – This is needed for organizations to access heterogeneous data sources. Infosys Next-Gen TDM provides a lightweight query engine that enables testers to mine lightweight copies that are protected.

4. **Data provisioning** – There are numerous challenges faced by testing teams in getting access to the right data. Large enterprises need approvals to access data from businesses and app owners. Infosys Next-Gen TDM provides an automated workflow for intelligent data provisioning. With this, testers can request data and manage entitlements as well as approvals through a simplified UX.

5. **Privacy-preserving synthetic data** – It is important to protect personal data residing in the data sources being curated for test data. There is always a risk of personal data being compromised when there is a large amount of training or testing data involved. It can result in giving too much access to sensitive information. Improper disclosure of such data can have adverse consequences for a data subject's private information. It may put data subject at more risk of stalking and harassment. Cybercriminals can also use data subject's bank details or credit card details to degrade subject's credit rating. Privacy-preserving synthetic data focuses on ensuring that the data is not compromised while maximizing the utility of the data. Differential privacy prevents linkage attacks, which cause records to be re-identified even after being anonymized for testing.

6. **Smart augmentation of contextual datasets** – Dynamic data can change its state during an application testing process. To generate dynamic data, the tester should be able to input the business rules and build both positive and negative test cases. Infosys Next-Gen TDM provides a configurable rules engine that generates test data dynamically and validates this against changing business rules.

7. **Image and audio file generation** – Infosys Next-Gen TDM can create audio files and image datasets for AR/VR testing using deep learning capabilities.

8. **Special file formats** – Customers need access to special communication formats such as JSON, XML, and SWIFT, or specific ones such as EDI files. Infosys Next-Gen TDM provides templates for generating various file formats.

9. **Intelligent automation** – Built-in connectors for scheduling the processes of data discovery, protection, and data generation allows testers to model, design, generate, and manage their own test datasets. These connectors include plug-ins to the CI/CD pipeline, which integrate data automation and test automation.

Accelerate through Next-Gen TDM Reference Architecture

As organizations look to deliver high-quality applications at minimum cost, they need a test data management (TDM) strategy that supports both waterfall and agile delivery models. With the rapid adoption of DevOps and increased focus on automation, there is also increasing demand for data privacy. Enterprises are fast moving from traditional TDM to modern TDM in order to meet the needs of the current development and testing landscape.

Infosys Next-Gen TDM focuses on increasing automation and improving the security of test data across cloud as well as on-premises data sources.
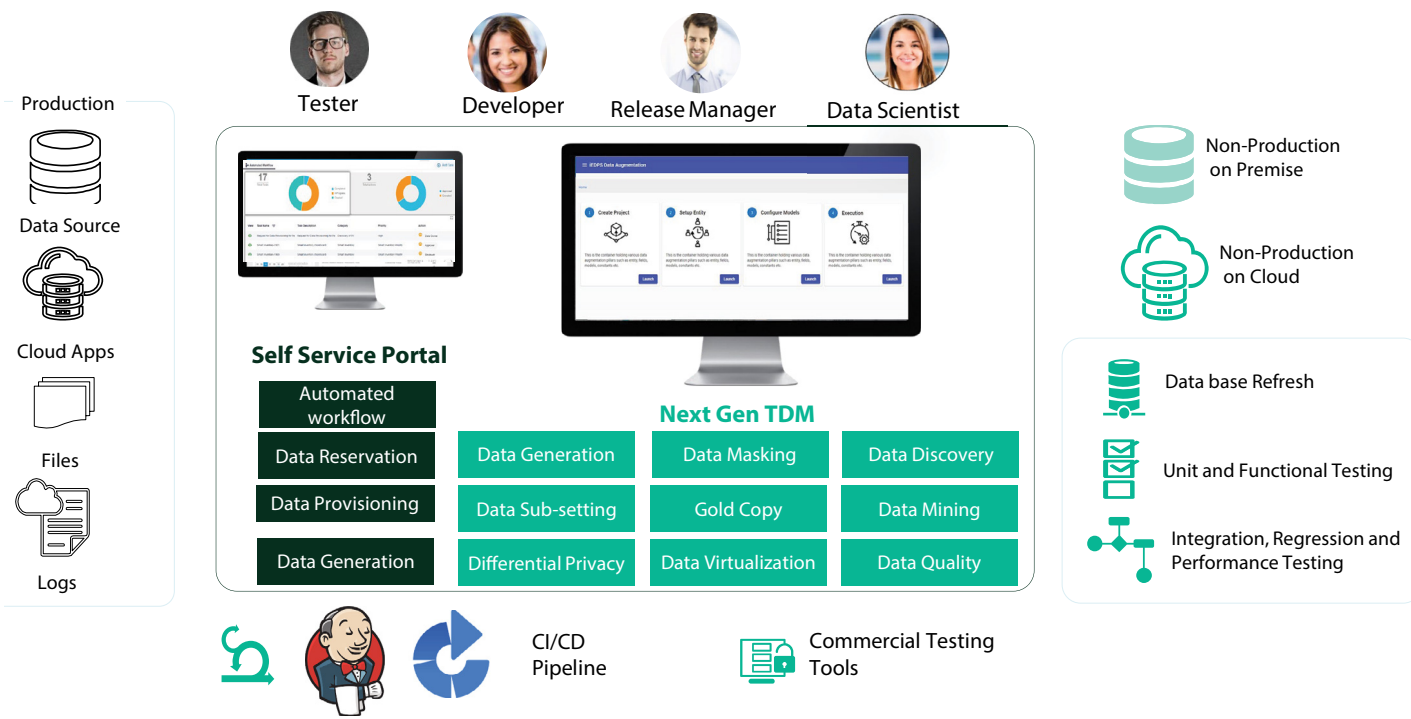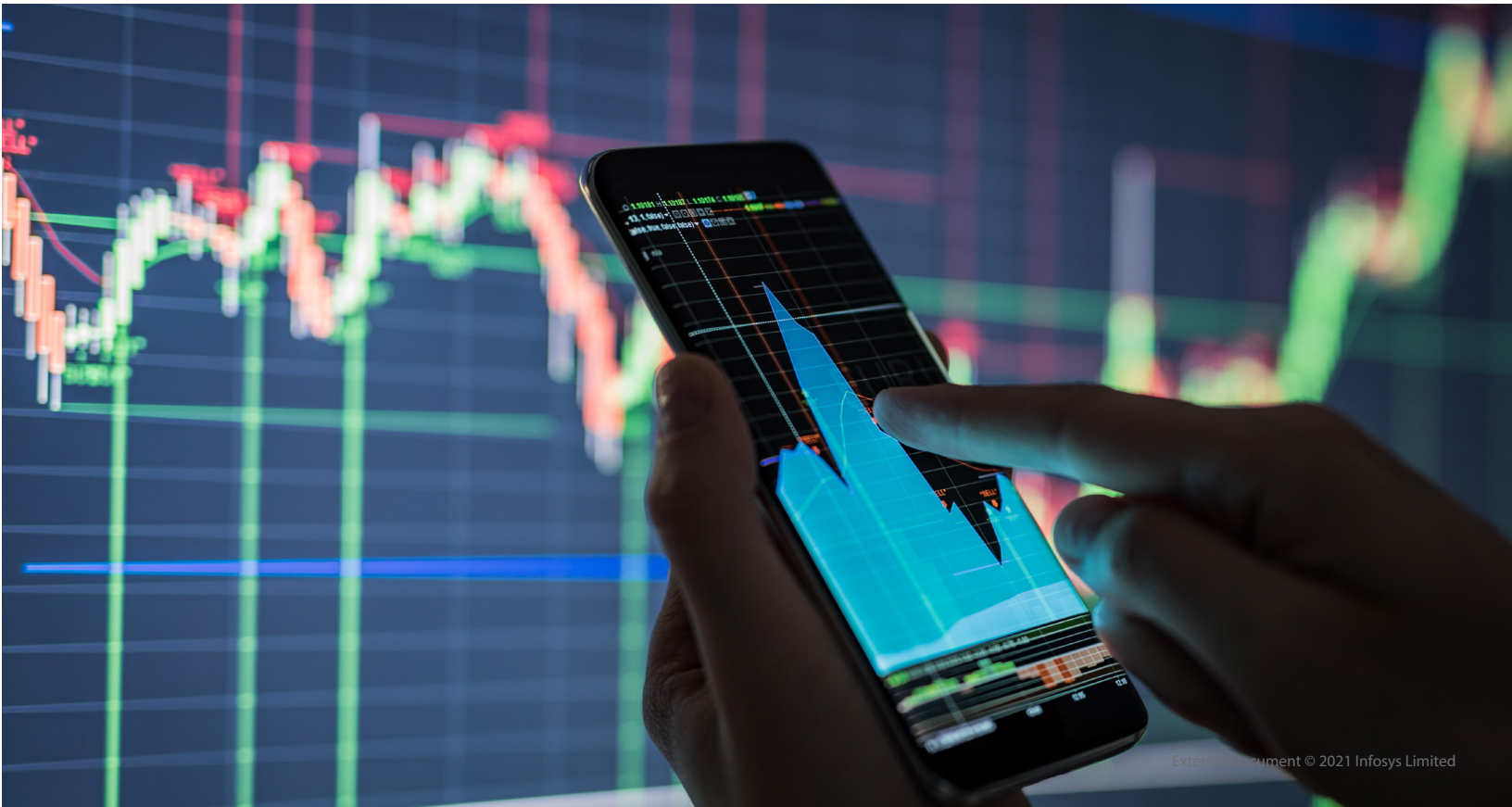

Figure 6. Infosys Next-Gen TDM reference architecture

The focus areas in digital transformation through this approach are:

1. **User experience** – Infosys Next-Gen TDM focuses on building specific data experiences for each persona, i.e., tester, release manager, developer, and data scientist. Its self-service capabilities offer simplified intent-driven design for better data provisioning and generation.

2. **Contextual test data generation** – There is a library of algorithms that helps teams generate different data types and formats including images, EDI files, and other unstructured data.

3. **Data protection for multiple data sources** – Infosys Next-Gen TDM connects to multiple data sources on cloud and on-premises. It provides a framework of reusable components for gold copy creation and sub-set gold copy. Data is masked and protected through a library of algorithms for various data types.

4. **Data augmentation** – The accuracy of AI and ML algorithms depends on the quality of training data and the scale of data used. The larger the volume and more diverse the training data used, the more accurate and robust the model will be. Infosys Next-Gen TDM generates high volumes of data based on a predefined data model, data attributes, and patterns of data variation for training, validating, and testing AL/ML algorithms.

5. **Integration through external tools** – To enable full-fledged DevSecOps, Infosys Next-Gen TDM has a library of adaptors that connect to the various orchestration tools in the automation pipeline.
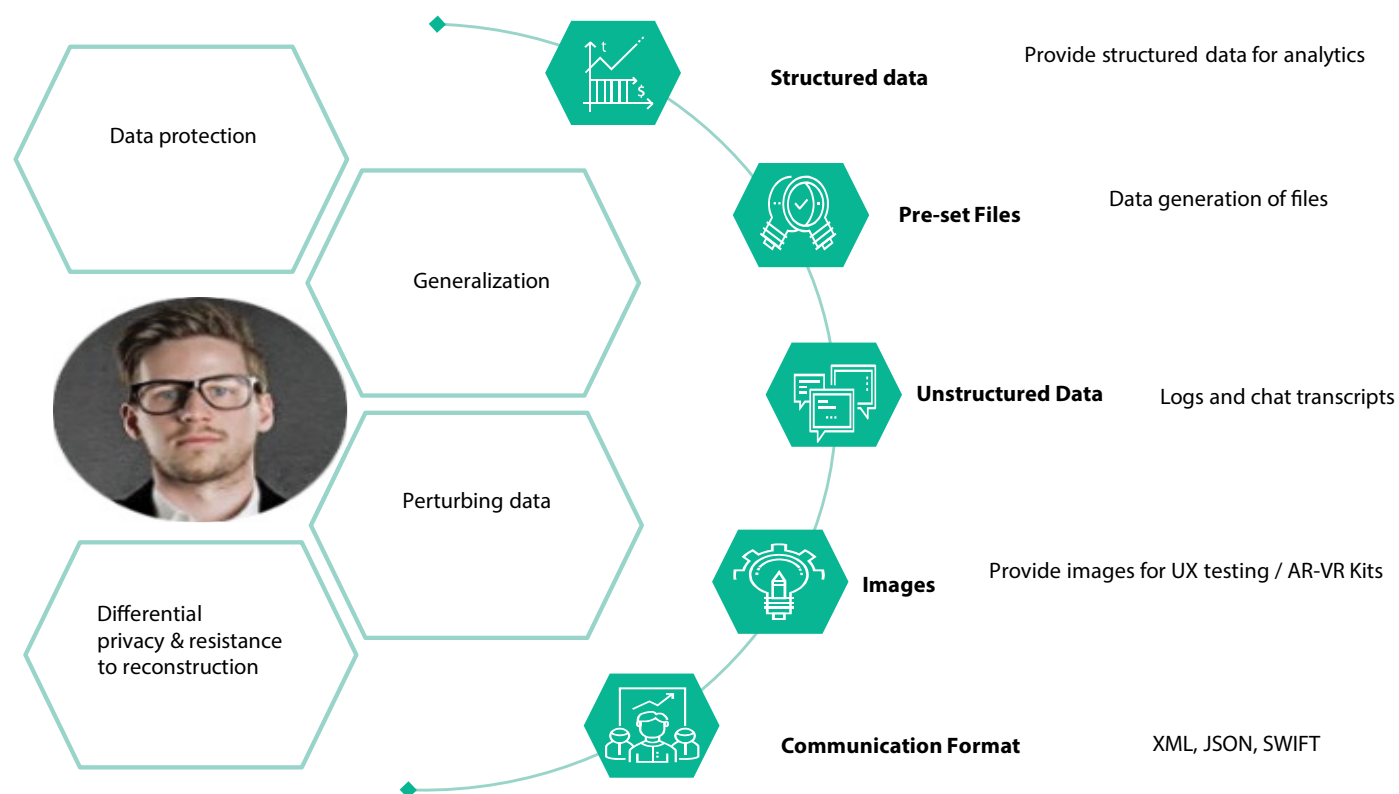

Structured data	Provide structured data for analytics	
Pre-set Files	Data generation of files	
Unstructured Data	Logs and chat transcripts	
Images	Provide images for UX testing / AR-VR Kits	
Communication Format	XML, JSON, SWIFT	

Figure 7. Contextual test data and its different formats

The Way Forward: Building Evolutionary Test Data for Your Enterprise

Production and synthetic test data can coexist in a testing environment, either to optimize their role in various testing operations or as part of a transition from one to the other. This may require the organization to think differently about test data and develop a roadmap for long-term continuous testing. To solve test data challenges, enterprises should focus on using evolutionary architecture to build contextual test data using a three-pronged strategy:

- **AI-assisted data prep:** Fitness functions – Focus on identifying the key dimensions of data that need to be generated for testing. Enhance feature engineering across multi-role teams to build the key fitness functions and models for data generation across each data domain and data type.

- **Focus on incremental change** – Help data architects focus on incremental change by defining each stage of test data management based on the tester's experience. This will enable testers to selectively pick the right data for different deployment pipelines running on different schedules. Partitioning test data around operational goals allows testers to track the health and operational metrics of the test data.

- **Immutable test data suite** – Focus on building an immutable test data environment with best-of-breed tools and in-house innovation to ensure the right tool choice for test data generation. This helps enterprises choose the tools best suited to their need, thereby optimizing total cost of ownership (TCO).

About the Authors

Avin Sharma

Consultant at Infosys Center for Emerging Technology Solutions (ICETS)

He is currently part of the product team of Infosys Enterprise Data Privacy Suite, Data for Digital ICETS. His focus includes product management, data privacy, and pre-sales.

Ajay Kumar Kachottil

Technology Architect at Infosys with over 13 years of experience in test data management and data validation services.

He has implemented multiple test data management solutions for various global financial leaders across geographies.

Karthik Nagarajan

Industry Principal Consultant at Infosys Center for Emerging Technology Solutions (ICETS).

He has more than 15 years of experience in customer experience solution architecture, product development, and business development. He currently works with the product team of Infosys Enterprise Data Privacy Suite, Data for Digital ICETS, on data privacy, data augmentation, and CX strategy.

Infosys®
Navigate your next

For more information, contact askus@infosys.com

Infosys.com | NYSE: INFY

Stay Connected