GenRocket

WHY DESIGN-DRIVEN SYNTHETIC DATA GENERATION SURPASSES STATISTICAL AND AI-BASED SYNTHETIC DATA APPROACHES

Executive Summary

Enterprise software engineering and quality assurance face an ongoing challenge: the lack of reliable, scalable, and compliant test data. Recent industry reports highlight that most enterprises struggle with insufficient, inconsistent, or privacy-compromised test data, hampering the effectiveness of software testing, automation, and overall quality engineering efforts.

While a number of vendors offer synthetic data solutions, some approaches are much better than others for Quality Engineering. **This paper outlines why a Design-Driven Synthetic Data Generation approach uniquely solves these challenges, in contrast to statistical replica solutions and AI-generated synthetic data from large language models (LLMs).**

The Test Data Problem in Enterprise Software Engineering

Recent studies identify key test data challenges facing enterprises today:

- Insufficient, irrelevant, or outdated test data.
- Data inconsistency and lack of referential integrity.
- Data privacy and compliance risk due to reliance on production data copies.
- Lack of control over data variety and negative test conditions.
- Inability to scale test data volume to meet performance or automation needs.

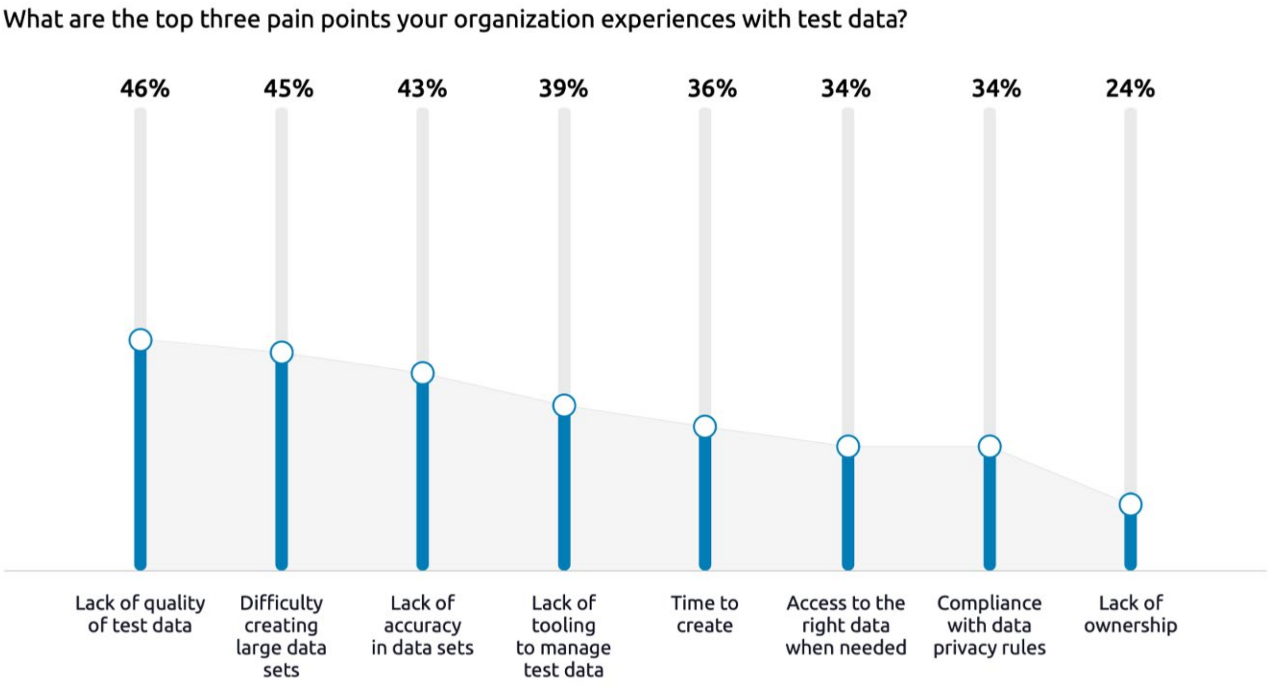
According to the *World Quality Report 2024*, 72% of QA teams now incorporate automation alongside manual testing, yet test data availability remains one of the top three constraints in achieving higher quality and faster release cycles.

QA teams are faced with test data challenges that ristrict their ability to provision quality test data at speed to support their automation efforts. The chart from the World Quality Report shown below illustrates the top pain points associated with test data provisioning for enterprise quality engineering organizations.

Challenges in creating test data

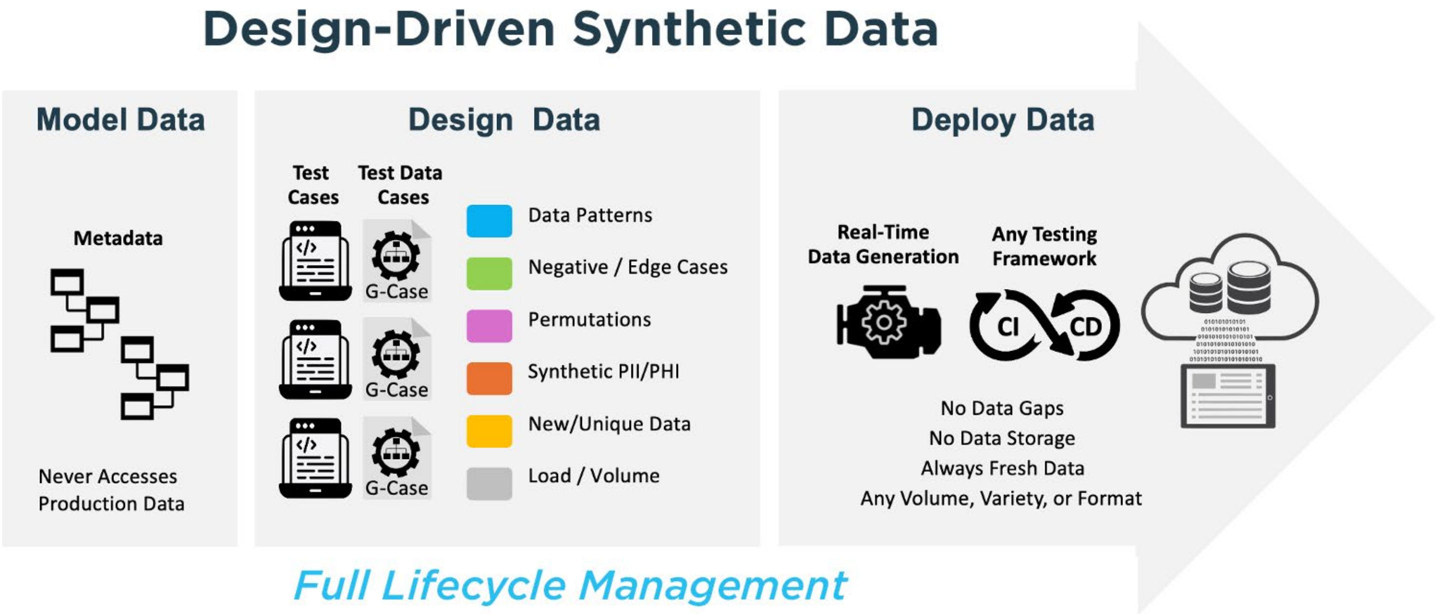
Fig 38 Top 3 test data pain points

WQR 2024 - Global Results



The Design-Driven Synthetic Data Advantage

Design-Driven Synthetic Data Generation is an advanced approach pioneered by GenRocket that allows development and testing teams to accurately engineer the test data they need to fulfill specific testing objectives. Unlike traditional test data solutions that rely on copying production data, statistically replicating datasets, or generating synthetic data using AI models, ***Design-Driven Synthetic Data*** **puts full control in the hands of testers. It empowers teams to define exactly what data patterns, conditions, relationships, and edge cases are needed** ensuring that every test case is supported by relevant, valid, and controlled data.



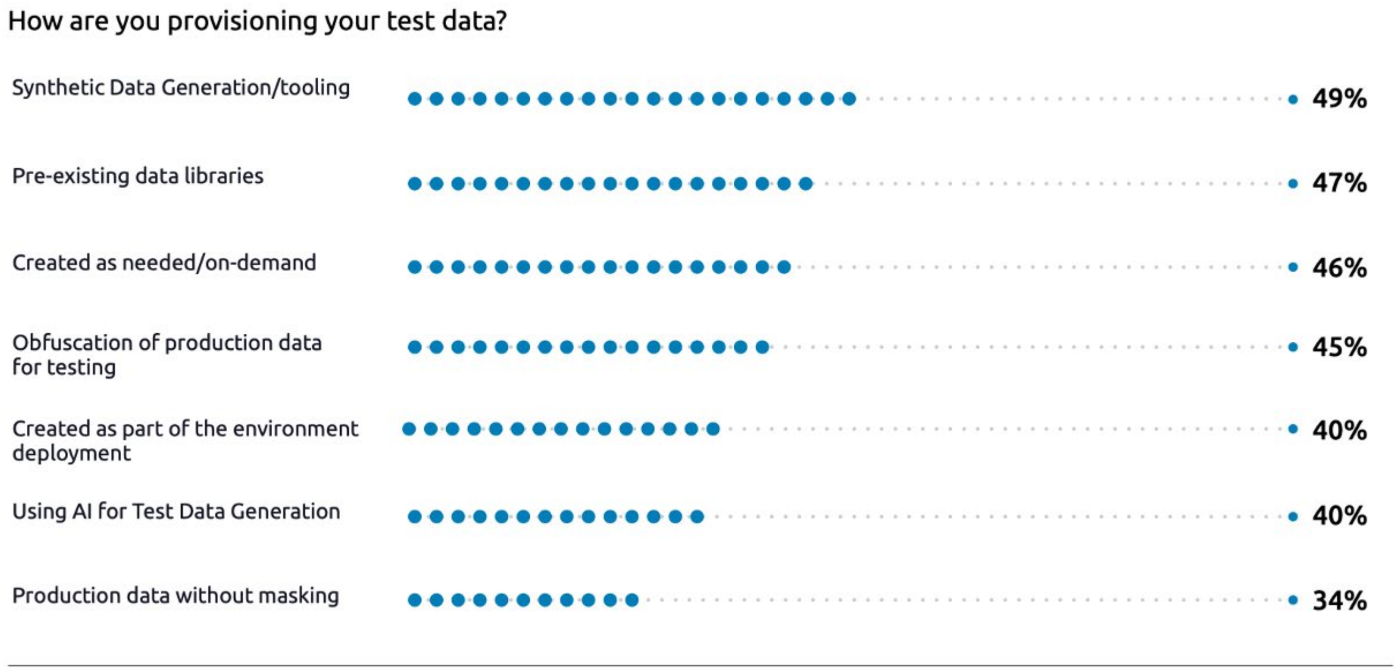
Key *Design-Driven Synthetic Data* benefits include:

- **Absolute data privacy:** No production data is required. Synthetic data is generated entirely from design models.
- **Full data coverage**: Edge cases, negative scenarios, boundary conditions, and complex relational data can all be generated intentionally.
- **Referential integrity:** Multi-table and parent-child data relationships are maintained by design.
- **Deterministic and repeatable:** Synthetic data sets can be generated and re-generated in the exact same volume, variety and format as needed.
- **Real-time and scalable:** Data can be provisioned on demand at enterprise scale.
- **CI/CD pipeline integration:** API-driven platform supports automated test data delivery in DevOps environments.

As revealed by the *World Quality Report*, synthetic data generation is emerging as an important alternative to the use of production data for testing.

AI and test data provisioning made possible

Fig 39 Approaches to test data provisioning WQR 2024 - **Global Results**



As organizations consider their transformation from production data to synthetic data, it’s important to understand the advantages and disadvantages associated with the different technologies and approaches that can be used for synthetic data generation.

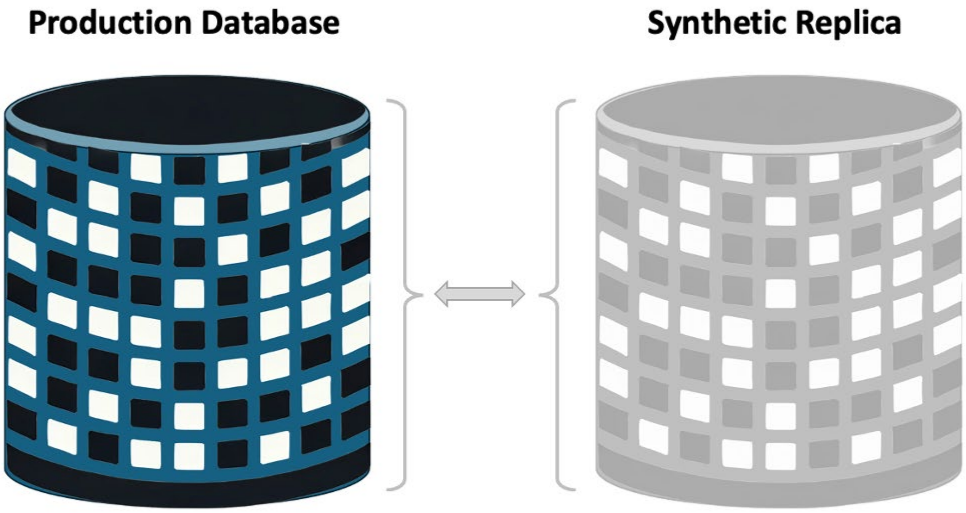
Limitations of Statistical Replica Synthetic Data

Statistical replica solutions generate synthetic data by analyzing production data and creating statistically similar records. Unlike the *Design-Driven Synthetic Data* approach describe above, these systems were not architected to produce data that does not already exist in production. While useful for secure data analysis, these solutions fall short of enterprise quality engineering requirements:

- They require access to production data to train statistical models, introducing compliance and privacy concerns.
- They lack control over data variety and edge cases.
- They replicate production data patterns, meaning gaps, errors, or missing conditions in production data are reproduced.
- They do not allow fine-grained control over data volume, structure, or scenario-based test data.
- Data refreshes require ongoing retraining and production data access.

Synthetic data generated based on a statistical profile of a production database inherit all of the same limitations associated with the legacy test data management paradigm.

Traditional/Synthetic TDM



Same Data Gaps

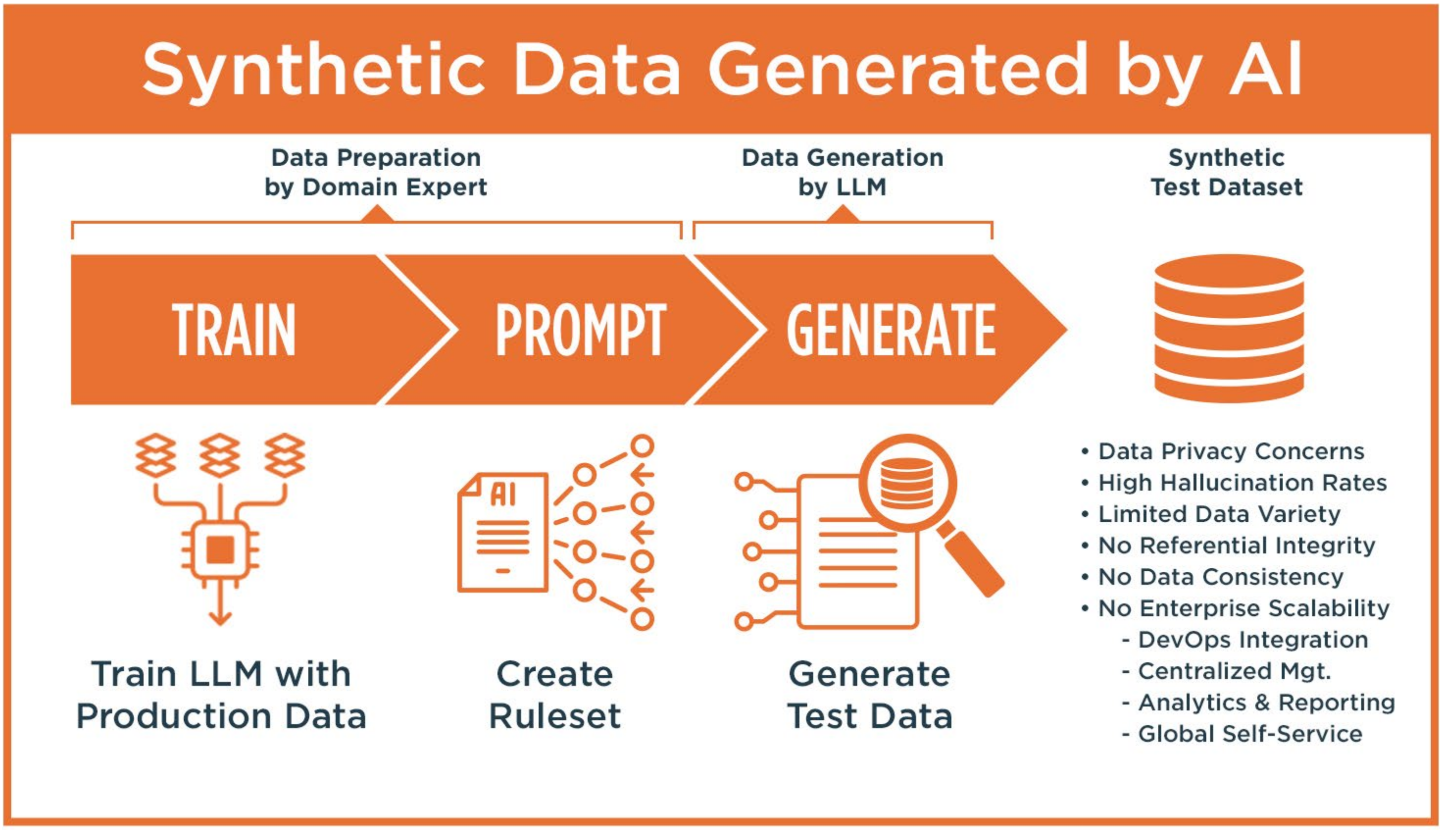
Must be Stored & Refreshed

Manually Augmented to Achieve Volume & Variety

Limitations of LLM-Based Synthetic Data

Recent interest in using GenAI and large language models (LLMs) to generate synthetic data has introduced new approaches but come with significant limitations:

- No referential integrity: LLM-generated data is unstructured and disconnected; complex data relationships cannot be maintained.
- Random and inconsistent: LLMs produce probabilistic outputs, not deterministic datasets.
- No control over edge cases: AI models struggle to reliably generate negative scenarios or specific boundary conditions.
- Low scalability: LLMs generate text-based data slowly and are not optimized for high-volume, structured data generation.
- Data privacy risks: LLMs may inadvertently reproduce real, sensitive information from their training data.
- Lack of integration: LLM outputs are not designed to integrate into automated CI/CD pipelines or enterprise systems.
- Hallucination risk: Studies show hallucination rates of up to 90% in some models (e.g., GPT-3.5 at 39.6%, GPT-4 at 28.6%, and Bard as high as 91.4%).



Synthetic Data Generated by AI

Data Preparation
by Domain Expert

Data Generation
by LLM

Synthetic
Test Dataset

TRAIN

PROMPT

GENERATE

Train LLM with
Production Data

Create
Ruleset

Generate
Test Data

- Data Privacy Concerns
- High Hallucination Rates
- Limited Data Variety
- No Referential Integrity
- No Data Consistency
- No Enterprise Scalability
  - DevOps Integration
  - Centralized Mgt.
  - Analytics & Reporting
  - Global Self-Service

Even when LLMs are trained on internal enterprise datasets, hallucinations persist, leading to inaccurate, misleading, or fabricated synthetic data that jeopardizes the reliability and accuracy of testing and compliance.

Conclusion

As enterprises modernize their software quality engineering processes, the limitations of statistical replica and AI-based synthetic data solutions become clear.

The *World Quality Report* highlights significant limitations in both statistical replica and AI-based synthetic data generation methods. Statistical replica solutions, which rely on analyzing production data to create statistically similar records, raise compliance concerns and lack flexibility, control over data variety, and the ability to generate edge cases or structured, scenario-based test data. They also require ongoing access to production data for retraining.

Similarly, large language model (LLM) based synthetic data suffers from major drawbacks, including the inability to maintain complex data relationships, inconsistent outputs, privacy risks, lack of scalability, and a high risk of generating inaccurate or fabricated data due to hallucination. The report concludes that only a *Design-Driven* approach can overcome these limitations, providing controlled, compliant, relational, and deterministic synthetic data at scale to meet modern software quality engineering needs.

GenRocket's *Design-Driven Synthetic Data* approach delivers:

- Complete control over data volume, variety, and structure.
- Compliance-safe synthetic data free from production dependencies.
- Full support for complex, relational data models and referential integrity.
- Deterministic, accurate, and hallucination-free synthetic data.
- The ability to provision test data on demand, at scale, as part of automated testing pipelines.

The GenRocket platform is purpose-built to meet the rigorous demands of modern software engineering with an enterprise-class synthetic data platform that maximizes data security, quality and operational efficiency.