WHITEPAPER

Test Data with AI

The XDM Agent-Based Solution

UBS HAINER

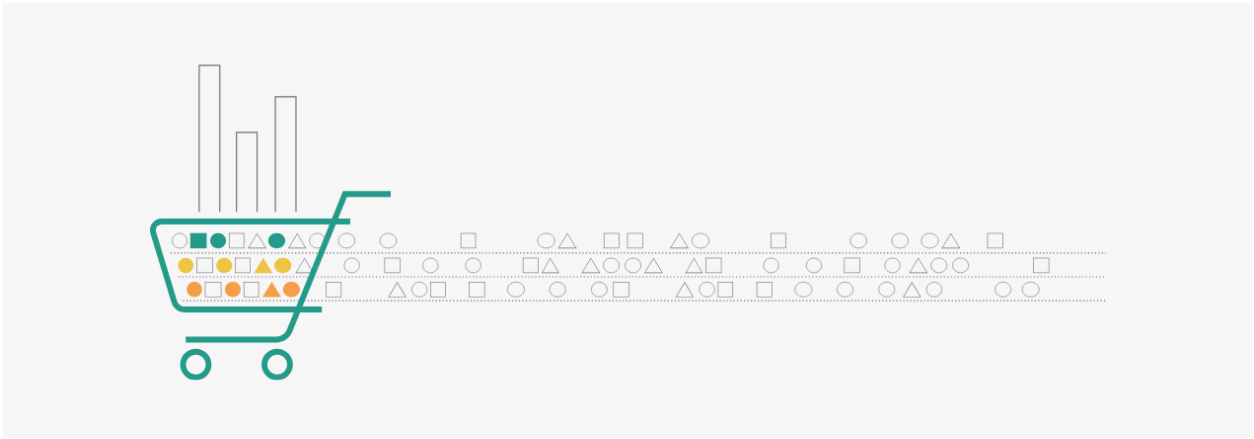
Limits and Leverage of AI in Test Data Provisioning

Agent-based Artificial Intelligence (AI) holds great potential for test automation – yet the foundation for any automation remains reproducible, consistent, realistic, and up-to-date test data. Can AI assist testers in creating the urgently needed data foundation within their test environments?

Artificial Intelligence is continuously reshaping the testing landscape – it promises to drastically reduce complexity and make automation effortless. As AI already accelerates software development, teams require fast and reliable access to test data for effective quality assurance. The timely provision of the right data to the right people is increasingly becoming a bottleneck. This is not just a productivity brake but, in the worst case, may compromise quality and security, as testers are forced to rely on outdated data or workarounds to achieve their goals.

Agent-based AI offers significant opportunities for test automation – but the bedrock of any automation is reproducible, consistent, realistic, and current test data. Can AI help testers establish this essential data foundation in their test environments?

Data Availability – An Underestimated Pain Point and a Cost Driver



Database administrators and operators face ever-increasing and accelerating demands, ultimately impacting productivity, quality, resilience, and compliance. A Gartner survey has shown that poor data quality now costs organizations an average of $12.9 million annually (https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality).

Test Data with AI

The XDM Agent-Based Solution

UBS HAINER

Creating data availability is anything but trivial, involving the simultaneous resolution of three challenges:

1. *Technical hurdles:* It must be ensured that referentially intact, realistic test data from various platforms or databases are provided.
2. *Compliance with data protection regulations:* Data privacy requirements must be strictly observed; access to production data must be strictly controlled and fully documented.
3. *Diverse requirements of data consumers:* Specific needs for test data vary considerably: a manual tester or developer may require precise data to test an individual function or reproduce production errors, whereas automation engineers often need reproducible mass data for nightly regression tests. These differing requirements across teams and departments present significant organizational challenges – especially in large enterprises or highly regulated industries.

AI Alone Cannot Solve the Test Data Availability Problem

It may seem tempting to "fight fire with fire," generating data on demand using AI rather than relying on masked production data. We see this approach – at best – as a partial solution:

Generative AI comes with inherent challenges, such as hallucinations, lack of reproducibility, and limited accuracy. These issues may be mitigated or even acceptable for certain use cases – for example, the ability to automate 80% of your tests through AI is a significant productivity boost, even if the remaining 20% must be handled manually. But what does this mean for test data generation?

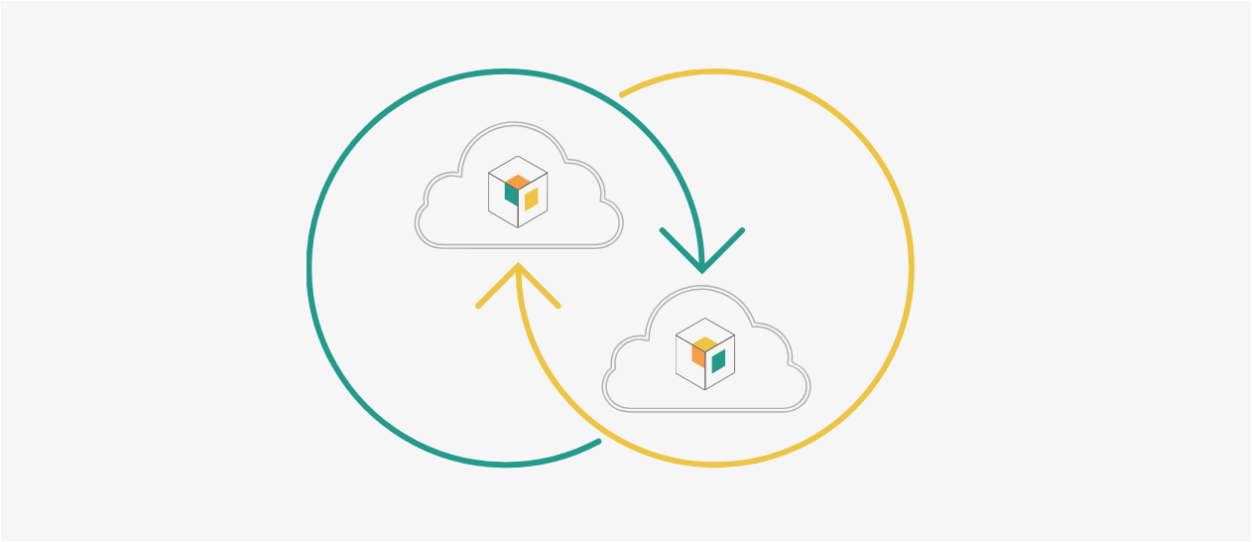
In the context of generative AI, an accuracy rate of 80% is often considered acceptable. But would you risk increasing your false positive rate by 20% simply because your test data are incorrect or inconsistent?

- *Even when testing is fully automated,* after each run it is necessary for a human tester to evaluate the failed tests – raising the original issue again: how can you ensure that data issues are not the root cause of test failures? How does the tester NOW obtain a referentially correct, realistic data set to validate this?
- *Lack of reproducibility* in test data can result in different scenarios failing in each test run. Such unreliability is intolerable for test data – while possibly acceptable elsewhere, in the context of testing, it is non-negotiable.

In the end, one cannot avoid production data or a well-planned data strategy, especially regarding edge cases and the resolution of production incidents. AI alone cannot replace the nuanced reality of production data – particularly for rare scenarios.

AI-driven data generation has its place in data augmentation, specifically in filling gaps or providing large volumes of data when production data is insufficient. Here, expert knowledge and practical experience are essential – both on the operator side and among data consumers.

From Reactive Process to Proactive Data Strategy



Many organizations are developing increasingly complex, homegrown solutions that require internal maintenance and quality assurance. This ties up highly skilled administrators (“operators”), whose primary responsibility is to keep the development pipeline running and sustain team productivity.

To reduce recurring operational tasks, more companies are choosing to implement off-the-

shelf data orchestration software.

+ In addition to saving effort and providing enhanced features, moving from a reactive approach to a proactive data strategy promises to greatly improve performance across all teams and departments. We often observe that after piloting such solutions in one area, other departments quickly follow, resulting in strong synergy effects.

We recommend a hybrid model in which realistic production data is supplemented by AI as needed and paired with targeted automation to drive productivity.

For shorter lead times and a better developer experience, an ideal data orchestration platform must:

- *Relieve highly skilled operators:* Enable them to focus on more strategically significant tasks.
- *Enable data-on-demand:* Data consumers have anytime access to required test data objects, even across different database platforms.
- *Enforce compliance and governance:* Role-based access controls and regulatory requirements are configurable within the platform; every distribution is fully auditable.
- *Offer flexible adaptability:* The tool can be configured to the specific needs of your organization and integrates seamlessly into existing CI/CD processes.
- *Leverage AI selectively:* Where production data is missing or unavailable, AI is used for data augmentation.
- *Configuration supported by generative AI:* For creating, evolving, and maintaining configurations, AI provides a useful and commonly employed aid.

The following section illustrates, using a real-world example, how successful implementation of a data orchestration tool can sustainably resolve technical, regulatory, and organizational challenges.

Use Case: Insurance Provider Optimizes and Scales Test Data Provisioning for Multiple QA Teams Across Diverse Database Technologies

Test Data with AI

The XDM Agent-Based Solution

An insurance provider operates in a highly competitive and tightly regulated market. With numerous application testing teams working across platforms such as Db2 z/OS, Db2 LUW, Oracle, and MS SQL Server, the organization faced the challenge of delivering robust, high-quality test data to support its continuously evolving applications.

Previously, test data was provided to the acceptance environment just twice a year. This irregular provisioning led to several issues:

- *Insufficient data quality:* Semi-annual delivery resulted in outdated or incomplete datasets, inadequate for comprehensive testing – leaving many potential issues undetected.
- *Misalignment among test teams:* Reaching consensus on the test datasets needed and their structure proved challenging among various test teams. These coordination issues led to inconsistent test practices and undermined the reliability of test results.

To overcome these challenges, the insurance company invested in an advanced data management tool. Following a rigorous proof-of-concept (POC) process evaluating several options, UBS Hainer's XDM emerged as the only tool meeting all requirements. Key decision factors included:

- *Scalable Data Processing:* XDM was able to copy 59 tables with up to 100 million rows from a DB2 environment, demonstrating its efficiency in handling large volumes.
- *Consistent Subset Extraction:* The tool reliably extracted a consistent subset of these tables – a critical capability for ensuring the test environment closely mirrors production.
- *Data Anonymization:* XDM offers robust, centrally integrated anonymization functionality,

ensuring sensitive production data is transformed in compliance with data privacy regulations.

Additional decision drivers were XDM's central control point, which streamlines the entire process through automation and repeatability, and outstanding support from UBS Hainer, which proved crucial during implementation.

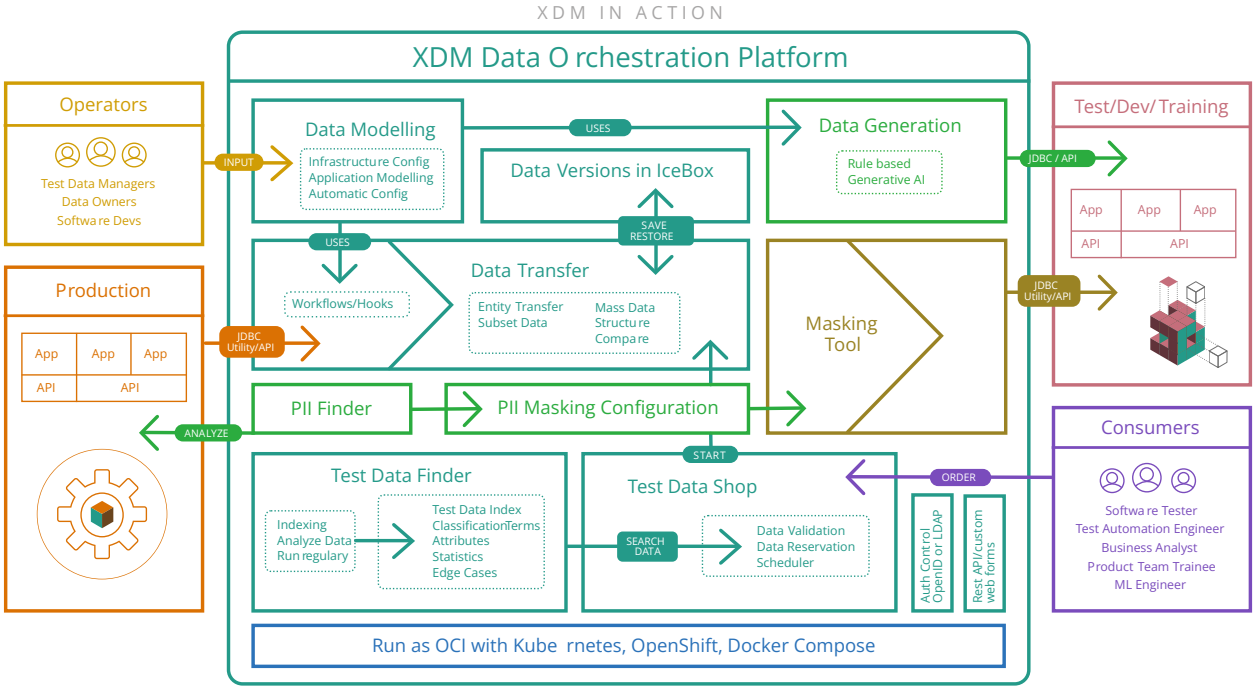
Deployment of the orchestration tool fundamentally transformed the organization:

- *Daily Test Data Provisioning:* Test data is now delivered daily from production, ensuring teams always have access to fresh, relevant, and up-to-date information. This dramatic increase in provisioning frequency has significantly accelerated test cycles and reduced time to defect detection.
- *Tailored Datasets:* The solution supports the individual needs of each testing team, so every group receives data sets precisely tailored to the features and functions under test.
- *AI for Data Augmentation:* In cases where formal correctness is not paramount (e.g., certain functional or load tests) or when insufficient production data is available (for new features), data is AI-generated. For example, fictitious "quotes" are generated to supplement existing policies.
- *Improved Operational Efficiency:* Automation and central management have significantly reduced manual operator workload, decreasing sources of error and enabling skilled personnel to focus on strategic tasks.
- *Enhanced Quality:* Transitioning to continuous data delivery has markedly improved overall test quality, operational resilience, and compliance adherence.
- *Pull Effect:* More departments are now planning to adopt the new technology.

Conclusion: Future-Proofing Through a Holistic Solution

Test Data with AI

The XDM Agent-Based Solution



By implementing a ready-made software solution with expert support for integration, the insurance company transformed its data provisioning approach from a slow and fragmented process to a robust, agile, and highly automated system that sets new standards across departmental boundaries.

This enables the company to maintain strict quality standards while meeting the demands of a rapidly changing regulatory and technological landscape. The combination of centralized control, efficient handling of large data volumes, consistent extraction, and effective anonymization not only fulfills current requirements but also positions the organization for future growth and innovation.