Artificial Intelligence & Machine Learning in Test Data Management

Praveen Bagare

bagare@gmail.com

Abstract

Artificial Intelligence (AI) and Machine learning (ML) are finding their roots in every stream of software development, be it self-driven cars, robo calls, meta humans or even digital clones. Test Data Management is no exception, significant use of ML for data discovery and profiling helps automate time consuming manual tasks into short, repeatable and effective outcomes. AI being leveraged to create Synthetic test data is just one example of how significant efficiencies can be brought into Test Data Management (TDM).

This paper will take you through key approaches to TDM solutions and how AI and ML can be applied across different stages of the Test Data Life Cycle (TDLC).

Biography

Praveen is a seasoned professional with over 20 years of IT experience in Program Management, Quality Assurance, and specialization in Test Data Management. He is passionate about solving complex challenges, driving innovation and excellence.

Currently, he is leading EPAM's Test Data Management Competency Center in the North America region and has successfully architected, implemented and Managed Testing and Test Data Management solutions across Fortune 100 companies in the USA.

He has authored a white paper on Test Data Management in Software Testing Life Cycle, is certified in multiple TDM tools, is ISTQB certified and is a senior member of IEEE.

1 Introduction

Machine learning has the ability to understand and learn from large volumes of data, and AI has been evolving to quickly bring back relevant information to the user faster and with more accuracy than ever before.

As AI is making leaps not only in autonomous cars, customer center calls, automated kiosks, and many other areas, it is also advancing in the test data management space. The ability to analyze data, learn, and re-create new sets is not just a concept anymore, but a reality today.

Let's take a deep dive into what AI and ML have to do with test data management.

2 What is Test Data Management?

TDM is the process of provisioning data for testing ensuring it is of high quality, provisioned in suitable quantity, right format, appropriate environment, and in the stipulated time. [1] It typically offers services such as subsetting, masking, reservation, synthetic generation, copying, cloning, reservation, ephemeral solutions and so on. Test Data Life Cycle (TDLC) is the process originating from a Test Data request to the stage where it is provisioned to the user. Under the hood the requests may be provisioned through one of the service techniques mentioned earlier.

3 What is Artificial Intelligence?

Artificial Intelligence (AI), was a term coined by Stanford Professor John McCarthy in 1955. It was defined by him as “the science and engineering of making intelligent machines”. [2] In other words AI is the technology that enables machines to perform tasks that typically require human intelligence.

4 What is Machine Learning?

Machine Learning (ML), is the process where systems learn from the data and improve through a feedback loop thus improving accuracy in every iteration. Typically, the larger the sets of source data, the more accurate the outputs are.

5 AI and ML in TDM Solutions

TDM solutions can be categorized into three broad approaches. Each of them have their own pros and cons and can be influenced by AI and ML.

5.1 Approach 1 - Conventional, leveraging Masking

These types of solutions focus on how production data can be masked and moved safely into lower environments for testing, thus helping with compliance to regulations like HIPAA (Health Insurance Portability and Accountability Act) [3], General Data Protection Regulation (GDPR) [4] and others. The process is not straightforward and can involve many complex steps:

- Determine if there is an appetite for a masking tool or will the process have to be done manually
- Identify and onboard the TDM tool that fits the masking needs of the organization (some factors to compare include databases supported, on-prem/cloud accessibility, security, support offerings and others)
- Find the personally identifiable information (PII), Personal Health Information (PHI), Payment Card Information (PCI) and other sensitive elements in the databases, files and other systems.

- Get this list of identified elements reviewed and approved by the application teams, security and other key stakeholders.
- The next step is to identify algorithms that will be used to mask these elements. The algorithms need to be signed off by security and compliance teams as well, to avoid reverse traceability.
- Post that, these masking algorithms need to be run on a copy of production data, probably in a staging environment or gold copy environment.
- After that we get into validating the masking by comparing the masked PII elements with production data. Once done, the data set is either subset or fully loaded into the target testing region and handed over to the application team for testing.
- A feedback loop is typically introduced where any slipped sensitive data can be identified, fixed and updated right back into the original masking rules.

As you can imagine, there are a lot of pros and cons to the solution as it can be time consuming and expensive to execute but can bring in the assurance of safety. The magic quadrant TDM tools in the market can cost an enterprise millions of dollars to procure and effectively implement these solutions.

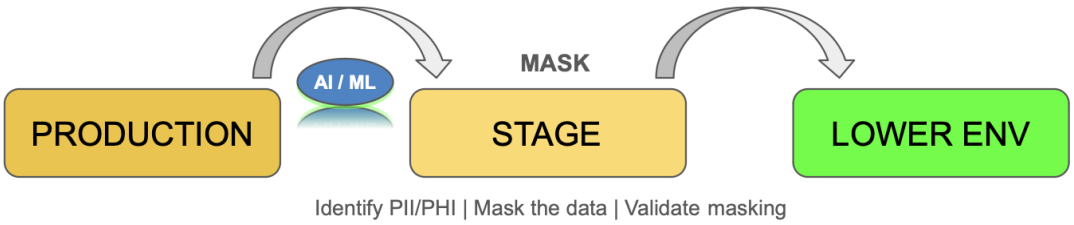


Figure 1 - Test Data Management leveraging masking

As seen in the image above AI/ML can play a part in identifying the PII/PHI elements and assigning appropriate generators/functions/algorithms to them. This speeds up the process and improves the efficiency significantly.

5.2 Approach 2 - New, leveraging synthetic data from scratch

This solution is relatively new and leverages the ability to examine the production database structure or the requirements and build synthetic data from scratch. These solutions are preferred when and where there are significant security concerns, such as in the healthcare and banking industries. Managing sensitive data becomes much easier because production data is not brought down to the lower environments. This may be a huge benefit when it comes to security and compliance team sign off. Typically, these solutions start by:

- Identify and onboard the right synthetic data generation TDM tool or tools.
- Understanding the table relationships in and across databases.
- Then adding further business rules into these functions or generators for data integrity.
- Assigning generators or functions to each column of data to be generated.
- Followed by generating sample test data sets and loading them to test environments.
- Validation of the data set through the application under test.
- Feedback loop to update the generators, functions and rules to create accurate data for testing.

The rules have to be well-defined, reviewed, approved, and tested. Commercially available tools can not only run these solutions on-prem but also in cloud or any other infrastructure. These tools allow the ability to write to the database directly or to files in different formats such as JSON, XML, and CSV, efficiently handling large volumes of data, potentially even millions of rows in seconds.

This solution is very effective in helping train machine learning models where the data is easily accessible to the learning system while addressing security concerns. Synthetic data may be a boon, but let's explore the other approaches too.

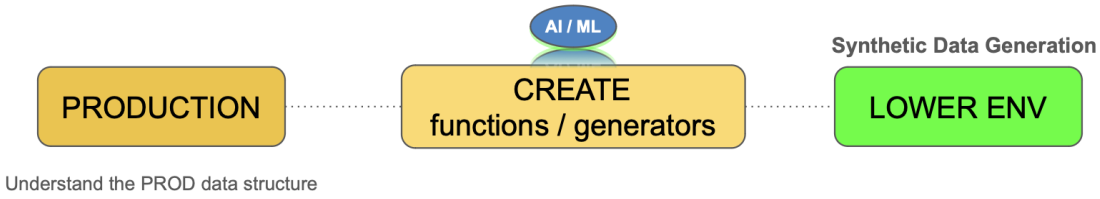


Figure 2 - Test Data Management leveraging synthetic data generation

As depicted in the above image AI and ML can play a key part in identifying and assigning the right functions or generators. This reduces human intervention, increases accuracy and speeds up the entire process.

5.3 Approach 3 - Futuristic, leveraging ML and AI to generate data

This approach examines production data and learns from it. The process can identify data types, the ratio of data distribution, and other mathematical entities. The machine learning process goes through the data and creates a model that can be used to regenerate similar data. Of course, there can be multiple iterations of learning, called epochs, which improve the model with every iteration. Once the model is available, security measures such as differential privacy can be applied. Once this model is secure, it can be moved to any lower environment and used to generate new sets of data. Here are the typical steps.

- Identify the subset of data on which you want to run the machine learning (biased, corrupt and irrelevant data rows can be eliminated for better performance).
- Configure the number of epochs (passes of training dataset through the algorithm), the relationships, pass through columns (e.g. policy number), custom requests (e.g. address) etc.
- Run machine learning (train) on the identified data set. This step can be Graphics Processing Unit (GPU) or Central Processing Unit (CPU) intensive.
- Move the differential privacy applied model to a lower environment where the data needs to be generated.
- Finally run the synthetic data generation (infer) to create desired volumes of data with similar attributes and distribution as the source data.

Figure 3 - Test Data Management leveraging machine learning and AI based generation

This solution comes with many benefits, including that the system automatically understands the source data through ML, which reduces human effort significantly. The generated data in the lower environment using AI has similar distribution as source, thus can be used for data analytics and business decision making. The data is not connected to production data, thus making it easier for Security and Compliance teams to sign off on these approaches.

However, there are some concerns, the generated data may not look very similar to the source data if the number of epochs are low. The process may have some hallucinations, and the accuracy may not be at 100%.

5.4 Comparison

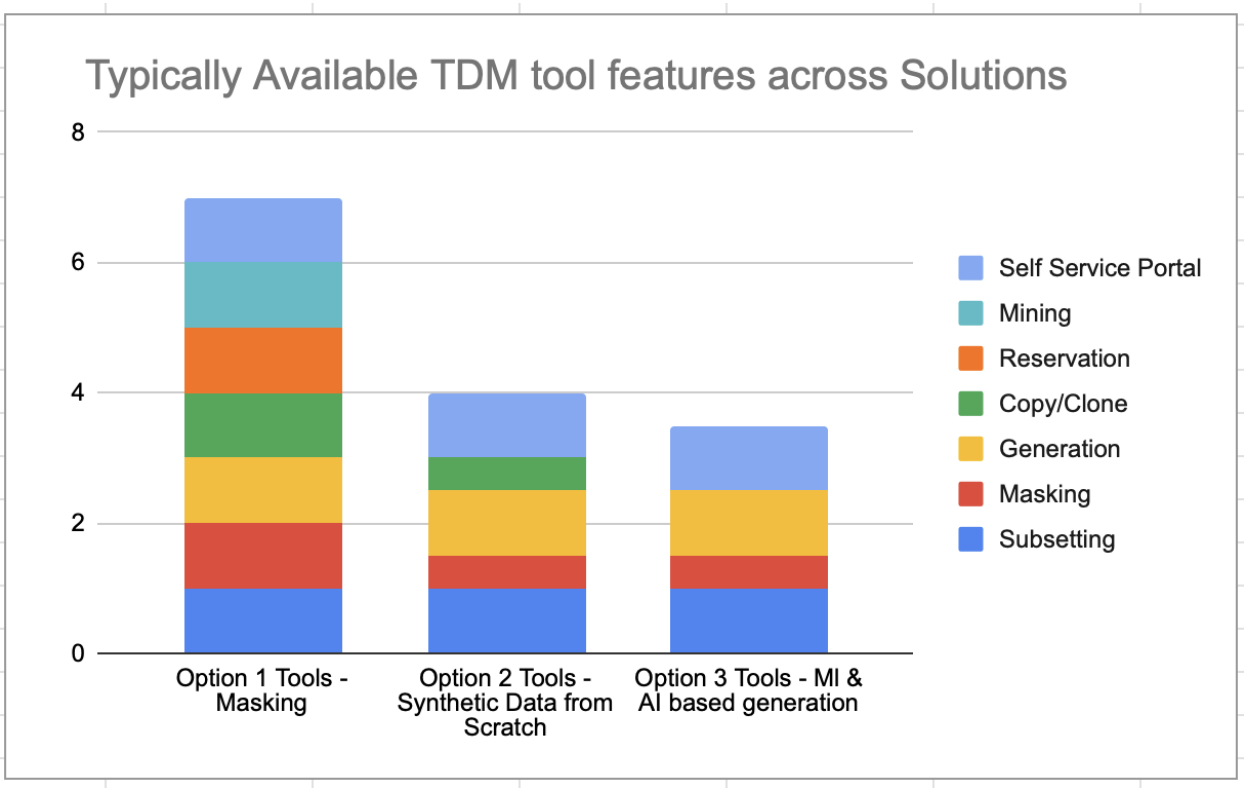
Let us compare the three TDM approaches side by side. Text highlighted in green are positive indicators.

	Approach 1 *(Masking)*	**Approach 2** *(Synthetic data from Scratch)*	**Approach 3** *(Synthetic data from sample data)*
Cost	High	Low	Medium
Implementation complexity	High	High	Medium
Maintenance	High	Medium	Low
Security/Privacy	Low	High	Medium
Ensure Data Integrity	Medium	High	Low
Return on Investment	Low	High	Medium
Effort	High	High	Medium
Time	High	Medium	Low

Table 1 - comparison of the three TDM approaches

6 TDM Tools

A wide variety of TDM tools are commercially available in the market for the approaches discussed above. Below are indicators as to how the different TDM services span across the tools.



Graph 1 - A graphical representation of TDM tool vs typically available features

A key point to note is that AI and ML tools may seem to have a limited coverage of services, but may be sufficiently able to solve the key asks from the client.

7 Case Study

Application under test: A simple two page web bank application for a user to submit an account opening form and be navigated to a confirmation page with a $200 account opening bonus.

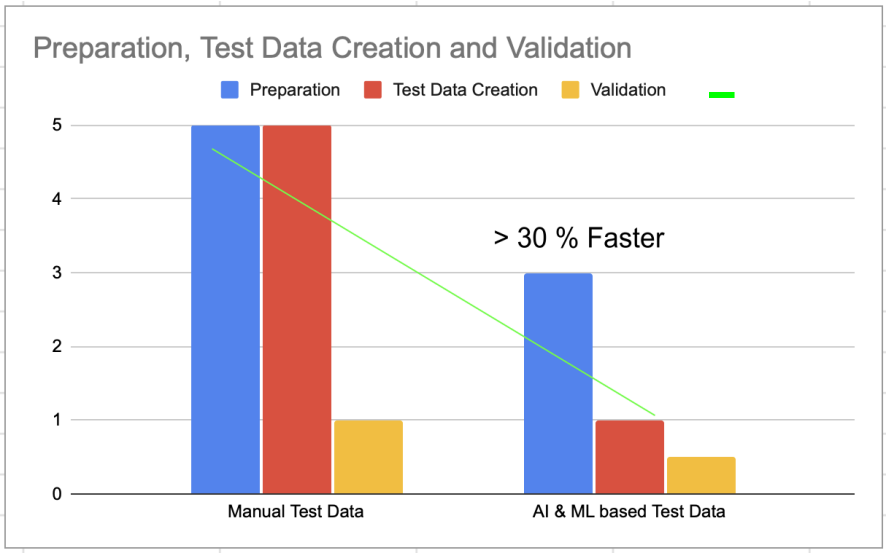
Database: A SQLite database with two simple tables, ‘Application’ and ‘AccountBalance’.

What was done? In this case study Approach 3 was used (Synthetic Data generation using an AI and ML based TDM tool) to generate the Test Data for testing. The ML solution was used to train a model on the two tables with a few thousand rows of data. The model was then used to infer/generate the test data with similar ratios as source, in the testing database. If test data had to be provisioned manually, it would have been extracted, masked and loaded into the target environment. It was observed that the manual test data preparation effort was over 30% more as compared to the AI/ML solution for testing this simple application.

This example gives you an indicator to the amount of effort and cost savings that effective TDM can bring to the table & further get enhanced with AI and ML solutions in TDM.

Benefits:

- Faster, thus saving time, effort and money.
- No Personally Identifiable Information (PII) or Personal Health Information (PHI) or Payment Card Information or other sensitive information flows down to a non-prod environment, thus reducing the risk of data theft.
- The source data distribution is maintained in the generated data, thus it can be used for data analytics and business decisions.
- In fact this type of data can be used for Machine Learning purposes as well.



Graph 2 - Effort savings of synthetic data generation using AI and ML vs Manual TDM

8 Conclusion

All said and done, even with advanced AI and ML, these approaches have their own positives and negatives. For a successful TDM implementation it is key that the TDM architect, application teams and other key stakeholders agree on a solution that best fits the needs of the enterprise but keep an open eye for the fast evolving TDM solutions in the AI space.

References

Tools:

1. https://app.grammarly.com was used to review grammatical correctness in some of the sections.

2. Google slides and docs were used to prepare the paper content.

Web Sites:

1. Praveen Bagare and Ruslan Desyatnikov. 2018. “Test Data Management in the Software Testing Life Cycle”. https://www.infosys.com/it-services/validation-solutions/white-papers/Documents/test-data-management-software.pdf (accessed Jun 01, 2024)

2. Christopher Manning. 2020. “Artificial Intelligence Definitions”. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf (accessed Jun 24, 2024)

3. “Health Insurance Portability and Accountability Act (HIPAA)” https://www.hhs.gov/programs/hipaa/index.html (accessed Jun 08, 2024)

4. “General Data Protection Regulation (GDPR)” https://gdpr.eu/what-is-gdpr/ (accessed Jun 08, 2024)