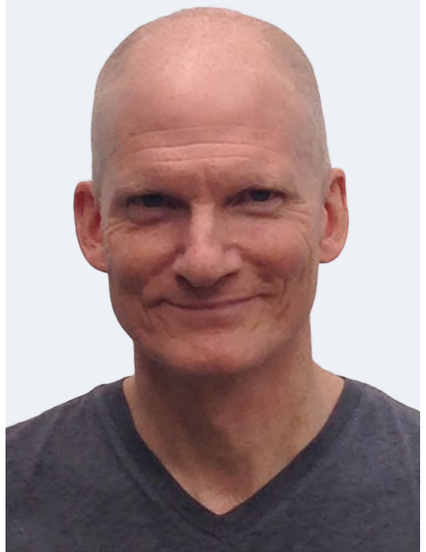
THE GREAT TEST DEBATE

Test data has become the center of attention for QA professionals looking to keep pace with the speed of development. But when should they use production vs. synthetic data for testing?

GARTH ROSE
CEO
GENROCKET

Garth runs a fast-growing software company based in Ojai, CA, which allows him to leverage 30 years of experience being a technology executive in software start-ups and publicly traded software companies



Test data provisioning has become a bottleneck that threatens the efficiency gains offered by new test automation technologies. As a result, test data represents a weak link in the chain for organisations implementing continuous integration and delivery.

Additionally, test data has been identified as a vulnerability for companies that must adhere to data privacy laws, like General Data Privacy Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), designed to prevent the accidental or intentional exposure of personally identifiable information.

To accelerate the speed of test data provisioning, meaningful change in the process and use of technology is needed; overcoming the threat of exposing sensitive customer information requires a fresh look at the sources of test data being used.

QA CHALLENGES

How can QA departments simultaneously maximise the speed, quality and privacy of test data while minimising the cost and complexity that come with provisioning it?

Companies are increasingly needing to address the challenge of keeping up with the accelerated pace of development as the bar simultaneously continues to rise for higher quality code and absolute data privacy. A sea change is underway in the form of synthetic test data that can be generated on-demand as an alternative to the traditional approach of subsetting and masking production test data.

But which approach is better? What are the trade-offs? How can IT professionals make the best decision for their environment?

These questions set the stage for a great debate about whether production test data or synthetic test data is a better solution for continuous testing. Here I'll introduce six essential test data criteria to serve as a basis for comparison between the two. Let's start by defining our terms more precisely.

PRODUCTION TEST DATA

Production test data is a copy of a production database that has been masked, or obfuscated, and subsetted to represent a portion of the database that is relevant to a test case. Production test data is frequently accompanied by a test data management (TDM) system to prepare, control and use the data. Commercial TDM systems can be expensive, costing upwards of hundreds of thousands of dollars for a typical enterprise deployment. Many organisations have chosen to develop their own in-house TDM systems and processes to save money and to provide a solution that more precisely meets their needs. TDM systems are typically accompanied by a highly controlled and centralised test data provisioning process.

SYNTHETIC TEST DATA

Synthetic test data does not use any actual data from the production database. It is artificial data based on the data model for that database. For the purpose of this article, we'll assume synthetic test data is generated automatically by a synthetic test data generation (TDG) engine. TDG engines generate synthetic test data on-demand and according to a test data scenario that represents the needs of a particular test case. Synthetic test data generation eliminates the need for traditional TDM functions, such as masking and subsetting, because test data can be generated on-demand and without sensitive customer information.

As a result, TDG systems can be decentralised and operate through a self-service model.

6 ESSENTIAL TEST DATA CRITERIA

There are six criteria often used to guide the decision between the use of production and synthetic test data. Each one is essential to the ultimate goal of eliminating the test data bottleneck and avoiding the risk of a data security breach. Each criterion is posed as a question, so you can ask yourself how each one applies to the needs of your organisation:

1. **Speed:** What are your time requirements for test data provisioning?
2. **Cost:** What is an acceptable cost to create, manage and archive test data?
3. **Quality:** What are the important factors to consider related to test data quality?
4. **Security:** What are the privacy implications of these two sources of test data?
5. **Simplicity:** Is it easy for testers to get the data they need for their tests?
6. **Versatility:** Can the test data be used by any testing tool or technology?

Let's consider each of these criteria one at a time. As you read them, consider your own test environment and how each criterion can have an impact on the efficiency of your operation.

SPEED: *What are your time requirements for test data provisioning?*

A recent survey of DevOps professionals described the provisioning environment as a "slow, manual and high touch process". In a survey of respondents from QA/testing, development and operations departments, they found that, on average, 3.5 days and 3.8 people were needed to fulfil a request for test data to support a test environment and for 20% of the respondents, the timeframe was over a week. The survey group used traditional production test data as their principle test data source.

What if this timeframe could be reduced from days to minutes? Synthetic test data that simulates real world data can be generated at a rate of 1000's of rows per second. Dynamically generated synthetic data eliminates the need to request production data from the TDM team and also removes the need to mask and subset the data for use by testers. With a decentralised self-service model, testers can provision their own data whenever they need it and simply discard the data when they have finished running their test.

COST: *What is an acceptable cost to create, manage and archive test data?*

Because production data must be prepared, managed and stored, the cost of provisioning the data must be burdened by the cost of a TDM system. This in turn leads to the purchase of a major TDM platform or the internal development and maintenance of a customised TDM solution. The cost can easily reach hundreds of thousands of dollars to procure, customise, support and maintain the platform.

If synthetic test data is being generated on demand, there is no longer a need for a TDM platform. Only the test data generation platform is needed with a complement of licences for the testers who need the ability to generate their own test data whenever they need it. This can lower the cost of provisioning test data by up to 90% when compared to a full-scale commercial TDM system.

QUALITY: *What are the important factors to consider related to test data quality?*

When provisioning production test data, the elements of data that must be managed include the age, accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little control over the quality of data that comes from production. With production test data, you only get what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data and edge case data. Testers are often forced to manually modify the production data into usable values for their tests. And some test data is too complex or time-consuming to build by hand so those data sets are never built.

Synthetic test data removes the guesswork that goes into creating a data subset. It is generated based on a test data scenario that specifies the nature of the data patterns and permutations required to cover all edge cases of the test. Further, the test data scenario is able to quickly generate data with a level of complexity that is almost impossible to do by hand.

Figure 1. shows a sample of the test data variations that can be specified by a synthetic test data scenario to support the needs of the test environment.

The table in Figure 2. provides examples of the synthetic test data output.

Another important data quality requirement is referential integrity – maintaining the parent child relationships between database tables that are represented by the test data. It is important for the synthetic test data generation engine to ensure referential integrity to preserve the consistency of the test data and the accuracy of the test results. The chart in Figure 3. illustrates the referential integrity concept with a variety of data tables that have parent, child and sibling relationships.

Figure 1.

Types of Test Data that can be generated (partial list)

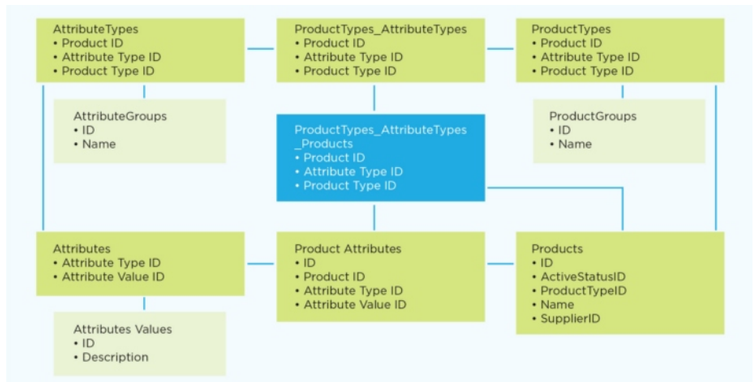
1. Patterned
2. Realistic
3. Sequential
4. Random
5. Edge Case / Negative
6. Null
7. Permutations
8. Percentages
9. Calculations
10. Images
11. Dates
12. Languages
13. Emoticons
14. Credit cards
15. Identification numbers
16. Addresses (in many countries)
17. Names (in many countries)
18. Health care formatted data
19. Financial services formatted data (e.g. credit card payments)

Figure 2.

Pattern	Realistic	Sequential	Random	Edge Case	Null
firstName1	Ms. Tereasa F. Saldana	001-01-0001	749-40-0182	749-40-0182	749-40-0182
firstName2	Mr. Everette Q. Groom II	001-01-0002	797-59-7445	797-59-7445	null
firstName3	Mr. Jules U. Hackney Jr.	001-01-0003	135-93-8060	135-93-8060	135-93-8060
firstName4	Mrs. Kristina J. Brick	001-01-0004	214-82-8447	214-82-8447	null
firstName5	Mr. Francisco M. Grimes II	001-01-0005	170-60-5224	170-60-5224	null
firstName6	Dr. Iona D. Starrett	001-01-0006	302-76-0978	302-76-0978	null
firstName7	Ms. Patricia O. Ingraham III	001-01-0007	266-20-5659	266-20-5659	266-20-5659
firstName8	Ms. Tracee M. Farah	001-01-0008	005-57-7667	005-57-7667	005-57-7667
firstName9	Mr. Alva I. Ziegler Jr.	001-01-0009	490-48-8084	490-48-8084	null
firstName10	Dr. Mike T. Youngblood II	001-01-0010	471-29-7519	471-29-7519	null

Data Variations for test data must cover every data scenario in order to discover defects for both expected and unexpected outcomes.

Figure 3.



test data?

In May 2018, the European Union enacted the GDPR, which requires any organisation doing business within the EU and the European Economic Area to provide data protection and privacy for all individuals. Failure to comply with GDPR carries heavy fines and penalties (up to 4% of global annual revenues) and it joins other security regulations in the United States such as HIPAA.

Test data provisioning must remove all PII, not only to be compliant with these laws, but to avoid subjecting the organisation to the enormously high cost of a data breach.

According to the Ponemon Institute , the cost of a data breach – including the costs of remediation, customer churn, and other losses – averages $3.8m (£2.9m).

Production test data relies on data masking techniques to obscure PII data, but no data masking process is perfect. And production data must still be handled by people during the masking process and archived on systems that can potentially be compromised.

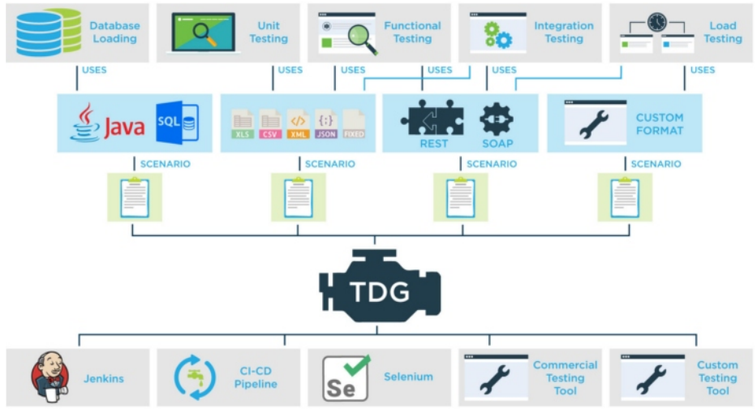
In contrast, synthetic test data is completely disconnected from production data, other than the data model used to generate it. This ensures a complete absence of PII from all test data and 100% compliance with all security regulations throughout the testing process.

SIMPLICITY: *Is it easy for testers to get the data they need for their tests?*

When compared to a $3.8m data breach, simplicity might seem like a trivial point of comparison. However, test data management systems that are complex and cumbersome introduce unnecessary delay into the provisioning process. As cited earlier, the average time to fulfil a request for test data to support a test environment is 3.5 days and 3.8 people. What is typically a centralised process for provisioning production test data perpetuates the siloed approach to development, testing and operations that DevOps is meant to eliminate.

Test data provisioning can and should be a simple, decentralised, self-service model that makes quality test data available to anyone at any time. This is the only way to eliminate the test data bottleneck and pave the way for continuous testing. Synthetic test data generation makes simple, decentralised test data provisioning possible with platforms that allow real-time test data to be created on-demand by anyone on the DevOps team.

Figure 4.



According to the *World Quality Report 2017-2018*, test environments and test data continue to be the "Achilles heel for QA and testing" and it was identified as the number one challenge in achieving the desired level of test automation by 48% of their survey respondents.

As the test data management market continues to grow – by an estimated 12.7% compound annual growth rate (CAGR), reaching $1bn in global revenues by 2022 – it also continues to evolve. The synthetic test data generation segment is expected to grow at the highest CAGR during this forecast period.

VERSATILITY: *Can the test data be used by any testing tool or technology?*

Versatility is another way of saying adaptability. The test data provisioning process should be adaptable to any testing environment, of any size, at any level of maturity, for any industry segment. That translates to integrating with a wide variety of frameworks and automation tools for seamless operations and supporting a variety of data formats for compatible test data output. It should also be capable of working with large databases with thousands of tables and potentially hundreds of different applications.

Test data management platforms tend to be database-centric, so when the test data use case is related to a database TDM's can usually satisfy the requirements, but often at a slower pace than is needed by continuous testing. Test data generation platforms have much more versatility so can satisfy a much wider variety of test data use cases and often the data is provisioned up to 10 times faster than TDM's due to the decentralised approach.

As you make your decision about production versus synthetic test data, be sure to closely examine the versatility of the platform.

The diagram in Figure 4. illustrates a test data generation platform integrating with a variety of frameworks and formats to maximise versatility.

MAKING THE TEST DATA DECISION

Consider the six essential test data criteria when making your own decision about the use of production data versus synthetic test data. Do they need to be mutually exclusive? Of course not. Production and synthetic test data can coexist in a testing environment, either to optimise their role in various testing operations or as part of a transition from one to the other. This may require you to think differently about test data as you develop a roadmap for your long-term continuous testing strategy. The idea is to be purposeful in your decision and to understand the implications on the speed, cost, quality, security, simplicity and versatility of your ultimate test data solution.