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?

**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 test engineers 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 fulfill 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.

**The Great Test Debate**

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.
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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.

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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.

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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.

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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.

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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.

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 burdensome due to 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 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 accuracy, variety and volume of data to be copied, masked and subsetted. Testers have little understanding of the quality of data that comes from production. When producing test data, one must understand what has been captured in the test data subset. Proper testing usually requires different permutations of data with negative test data scenarios that specify the nature of the data patterns and permutations. Test data provisioning must remove data variations that can be specified by the test data generation engine to ensure referential integrity – the test data and the accuracy of the test data 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.