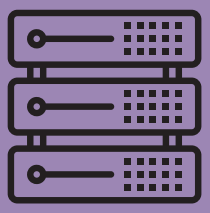
ACCELARIO

Synthetic Data

ACCELARIO

Test data is key to Quality Assurance testers focused on keeping up with the increased delivery velocity facilitated by Agile methodologies, Dev/Ops, & Data/Ops practices. But when should production vs. synthetic data be used for software testing?



Test data provisioning has become a bottleneck that threatens today's orchestration and automation technologies' efficiency gains; its inclusion is vital for any CI/CD (Continuous Integration/Continuous Delivery) system.

However, test data can be a vulnerability for companies that must adhere to data privacy laws. This vulnerability arises when Personally Identifiable Information (PII) and other sensitive data is exported or copied into lower environments that aren't as secure as the production environment. Privacy laws include the General Data Privacy Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), which were designed to prevent accidental or intentional exposure of PII.

Meaningful change in the process and use of technology accelerates test data provisioning. A fresh approach to provisioning test data to developers and Quality Assurance teams can prevent PII from being exposed in lower environments.

Synthetic data must be created manually or programmatically. Manual data creation is problematic because it is tedious and typically produces insufficient data sets or randomization inadequacies (in their relation to the data types) that reduce system/code testing accuracy. Synthetic data can also be created programmatically via a script, but again, the data and the script are only as good as the individual who created it.

The optimal solution is an automated method to create synthetic data using the production application or database data model. The automation captures the production data model and populates it with the fictitious records based on user-defined parameters.

ACCELARIO

Uses for Synthetic Data

Using synthetic or manufactured data in development and testing environments delivers the following advantages:

Enables testing earlier in the development cycle, even before there is any production data available	Eliminates the need for anonymization as it is devoid of PII	Supports AI (Artificial Intelligence) system training
Delivers a larger dataset than production, if production even exists	Provides QA testing with greater statistical accuracy than production data	Provisions data for the development of applications that interact with SaaS systems like Workday, SFDC, Netflix, etc... when accessing production data from the lower environments is not an option.

Benefits of Synthetic

Generating data that mimics the real thing may seem like a way to create infinite testing and development scenarios. While there is much truth to this, it is essential to remember that any synthetic models derived from data can only replicate specific properties of that data, meaning that it'll ultimately only simulate general trends.

However, synthetic data has several benefits over real data:

Overcoming privacy restrictions: Real data may have usage constraints due to privacy rules or other regulations. Synthetic data replicates the real data's critical statistical properties without exposing any PII, eliminating the issue.

Simulating conditions not yet encountered: When real data does not exist, synthetic data is the only solution.

Immunity to some common statistical problems: These include item nonresponse, skip patterns, and other logical constraints such as negative use cases.

Focuses on relationships: Synthetic data aims to preserve the multivariate relationships between variables instead of specific statistics alone.

The use of synthetic data will increase as data becomes more complex and more closely guarded within the enterprise.

Accelario Synthetic Data

Accelario Synthetic Data generation allows end-users to create synthetic data through a self-service Continuous Data Platform portal that connects to both the source and target databases. The platform provides a list of importable data models or schemas, which are used as the basis for the synthetic data. You can limit the number of rows in the generated tables based on a greater than parameter and a reduction factor that you provide. You can edit the data model on which the synthetic data will be based on. This is useful for situations when you add a column or make other changes in your test or development version of the applications.

Accelario can also create the needed schema for the synthetic data using the editor. This method is helpful for situations where the database structure is needed before the production database is created. In addition, the editor allows for the selection of rules and parameters for each column that determine how the data is generated. Rules categories range from Basic Rules such as randomization of numbers and strings, Personal Rules governing last names, addresses, and credit card numbers, along with Advanced Rules that allow fields to be filled from lookup lists, templates, and formulas.

Once configured, data generation is as simple as clicking ‘Run’ and selecting the target database in the development or test environment. Synthetic data generation is monitored in the Accelario Continuous Data Platform user interface.

Conclusion

Even with the advancement of production data anonymization, there is still a need for synthetic data in development and test environments. Synthetic data is not just valuable for situations when production data is unavailable. It is also useful when statistical distribution accuracy is needed in the data or for training AI systems. Synthetic data has the added benefit of not containing PII.

With Accelario's Continuous Data Platform and Synthetic Data Generation self-service model, developers and testing professionals control the process of generating synthetic data whether they import the schema from a production database or create one before a production database is available. By using Accelario Synthetic Data Generation it's possible to limit or expand the number of records generated, define rules used to populate the target database, and monitor the data generation. The whole process can be completed without the need for a DBA or IT operation staff involvement.

Sources

Why synthetic data is about to become a major competitive advantage by Gautier Krings riaker.com

The Ultimate Guide to Synthetic Data AI Multiple