GenRocket

USE CASE

AN INNOVATIVE APPROACH TO DATA SUBSETTING AND MASKING

What is Data Subsetting?

Data subsetting is the process of selecting a representative subset of data from a larger dataset. Subsetting provides a smaller, manageable volume of data that reduces storage costs and speeds up the testing process. By selecting a subset that is most relevant to the test scenarios, you can ensure that the tests are more meaningful and aligned with real-world use cases.

What is Data Masking?

Data masking is the process of disguising original data to protect sensitive information while maintaining the data's authenticity and usability. In many industries, especially those subject to strict data protection regulations (like finance and healthcare), it's essential to ensure that sensitive data doesn't fall into the wrong hands. Data protection regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), mandate that personal data be protected.

The GenRocket Solution - A "Best of Both Worlds" Strategy

In today's data-driven world, where the quality of software is paramount and the sensitivity of data is a primary concern, ***GenRocket provides an optimal blend of production data and synthetic data for more effective testing.***

While many quality engineering organizations are shifting towards synthetic data for testing, the vast majority of dev and test teams still lean on masked production data. Recognizing this pattern, GenRocket offers a solution that melds both worlds:

Intelligent Data Subsetting

The GenRocket solution allows teams to efficiently query and extract meaningful subsets from production databases for on-the-spot testing. This system extracts data from source production databases, either encompassing all tables or specific related data tables.

GenRocket's intelligent data subsetting solution allows:

- Subsetting based on defined values, percentages, or specific numbers of rows.
- Maintaining referential integrity even across different schemas.
- Reducing vast databases into manageable data subsets. For instance, transactions from all 50 states can be filtered to represent just one or two, all while ensuring referential integrity.

An impressive internal benchmark showcased the efficiency of GenRocket's solution. In one scenario, it took less than two minutes to provision a subset containing 8 million records.

Synthetic Data Masking:

Here, GenRocket truly stands out, allowing the dynamic masking of sensitive database fields using real-time synthetic data replacement.

Synthetic data masking capabilities include:

- Field-level masking for popular SQL databases and a variety of file formats.
- Real-time replacement of sensitive data elements with 100% secure synthetic data
- Assured complete compliance with all global data privacy regulations.
- Total security as only metadata is accessed; sensitive data remains untouched.
- Rapid and efficient data masking compared to conventional techniques.
- No need for data storage or reservation; fresh synthetic data can be generated for every test run.

GenRocket supports a wide variety of databases, including Oracle, MS SQL Server, IBM DB2, PostgreSQL, and MySQL. And supports file masking for numerous file formats such as CSV, JSON, XML, X12 EDI and more.

Combined Strength: Intelligent Data Subsetting & Synthetic Data Masking

When combined, GenRocket's data subsetting and masking capabilities form a comprehensive solution that drastically accelerates the time to provision data for testing. The GenRocket platform ensures test teams always have fresh data sets at their disposal, negating the need for data reservation or refresh.

Furthermore, GenRocket enhances its offering with the ability to augment data subsets with additional synthetic data. ***Synthetic Data Augmentation enables teams to supplement masked production data with controlled and conditioned synthetic data to deliver a more comprehensive testing dataset.*** This enriched dataset enables both positive and negative data testing, edge case conditions, and scalable data volumes for load and performance testing.

Simplified and Efficient Test Data Management

The GenRocket platform offers a highly streamlined process for test data provisioning. Some highlighted features include:

- Using XTS (Extract Table Schema) to seamlessly import data models while maintaining referential integrity.
- Visually showing the data model for ease of navigation and selection.
- Micro-subsetting that allows dev and test teams to craft smaller, test-case specific subsets and store them in a library of subsets and micro-subsets .
- High-speed data delivery, delivering about 5 million rows per minute to ensure on-demand data delivery for each test case in a matter of seconds.
- Enhanced security through synthetic data masking, a high-speed synthetic data replacement process that outperforms traditional obfuscation, which tends to be slow and which uses an algorithmic approach that can be reverse engineered.

Key Take-Aways

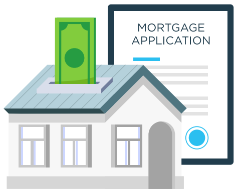
In the ever-evolving landscape of software testing, GenRocket's Test Data Automation solution offers an innovative and advanced alternative to traditional TDM. ***By providing an integrated platform for synthetic data generation, subsetting, and masking, GenRocket ensures dev and test teams have optimal datasets, tailored to their specific needs, and available on-demand through a self-service platform.***

Data Subsetting and Masking in Financial Services

The financial services sector deals with a significant amount of sensitive data, making it a prime industry for the application of data subsetting and masking. Here are some practical use cases.

Credit Card Transaction Analysis

When testing a new fraud detection algorithm, a bank uses data subsetting to extract a reduced set of transactions from a large production database. They select transactions that have been flagged as suspicious in the past year. They also ensure personal information, such as customer names, card numbers, and CVVs, is replaced with synthetic data to ensure data privacy.



Mortgage Application Processing

A financial institution is developing a new mortgage application system. They subset data to extract a representative sample of mortgage applications from the past six months. All personal details, including names, addresses, social security numbers, and financial details, are synthetically replaced.

Mobile Banking App Development

A bank is launching new features in its mobile app. They extract a subset of user data, including login habits, usage patterns, and transaction types, to simulate real-world testing scenarios. Usernames, passwords, account numbers, and transaction details are all replaced with synthetic data to ensure real user data remains confidential.

High-Frequency Trading Systems

A trading firm is improving its high-frequency trading algorithms. They subset trading data from specific volatile days to test the robustness of their new algorithms. The identities of traders, their specific trading patterns, and trade sizes are synthetically replaced to protect individual and institutional trader information.

Regulatory Compliance and Reporting

Financial institutions often provide data subsets to regulatory bodies for compliance checks. Instead of handing over entire databases, they can extract specific sets relevant to the regulatory requirement. To ensure confidentiality, financial institutions replace sensitive customer data with synthetic alternatives before sharing datasets with regulators.

Data Subsetting and Masking in Healthcare

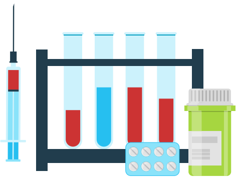
The healthcare industry, with its vast amount of sensitive patient data, has a critical need for data subsetting and masking. Here are some practical use cases in healthcare:

Electronic Health Records (EHR) System Implementation

A hospital is migrating to a new EHR system. For testing, they extract a subset of health records, specifically those with complex medical histories, to assess the new system's efficiency. The EHRs chosen for testing have sensitive patient information such as names, addresses, and medical history. These details are replaced with synthetic data to ensure patient confidentiality.

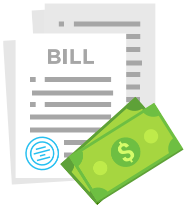
Insurance Claims Processing

An insurance company is developing a new algorithm to detect fraudulent claims. They subset data to extract a mixture of claims: those flagged as suspicious in the past, high-value claims, and routine smaller claims. Personal details, medical records, policy numbers, and payment details are replaced with synthetic data, allowing for realistic yet secure testing.



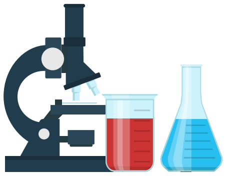
Pharmaceutical Research and Clinical Trials

A pharmaceutical company is conducting analysis on drug reactions. They subset data to extract records of patients who took a specific combination of medications. All patient identifiers and personal data are replaced with synthetic information, ensuring the anonymity of participants while preserving the integrity of the study.



Hospital Billing and Payment Systems

A hospital is upgrading its billing and payment system. They extract a subset of billing data that includes unpaid bills, bills with disputes, and routine paid bills. Patient names, treatment details, policy numbers, and payment details are replaced with synthetic data for secure testing.



Health Analytics and Predictive Modelling

A research institution is building a model to predict disease outbreaks. They subset data from hospitals in regions with recent outbreaks. All patient-specific information and medical histories are replaced with synthetic data to ensure confidentiality during analytical modeling.

Visit Our Knowledge Base

To learn more about GenRocket's Intelligent Data Subsetting and Synthetic Data Masking solution, visit our Knowledge Base and read about the G-Migration+, the GenRocket component that is used to perform subsetting and masking operations. You can also view an informative solutions video that explains how it works.

VIEW THE KNOWLEDGE BASE ARTICLE