PRIVITAR

The 6 principles of privacy-safe Test and Dev data

How advanced privacy techniques preserve the value of your data while taking the risk out of Testing and Development

There's still a persistent notion that only raw, sensitive data will do for successful Test and Dev. It's based on two factors that have traditionally made it hard to readily provision useful data to Test and Dev environments:

- Synthetic data isn't good enough when it comes to referential integrity and data complexity
- Masked data is only provided after a slow, manual process – which rarely goes far enough in eliminating risk (we've written a blog about it).

The good news: these things were true yesterday. New, sophisticated privacy techniques make it possible to achieve levels of data utility that have, so far, only been present in raw data.

Here's a list of 6 characteristics of valuable Test and Dev data

...and how a modern approach to privacy saves time and retains value, while protecting sensitive information.

1. It preserves the original data types and formats

The more closely your Dev and Test data resembles your production data, the more you can trust your testing. Advanced privacy techniques make sure that your masking or tokenisation processes are format-preserving, e.g. the tokenised form of an email address is still in the format of a valid email address.

2. It preserves the referential integrity of the data

Modern masking techniques ensure that the process is consistent, meaning that for each occurrence of the same input value, it produces the same masked output. This applies to the rows of a single table, or across multiple tables. It also applies over time, when tokens are re-used and known values come up again when new datasets are available. When the consistency of key values is preserved, advanced data privacy techniques will maintain referential integrity among tables and files. This applies both to individual datasets and across all your data at scale, if needed.

3. It preserves data complexity

Referential integrity is hugely important to do meaningful, representative testing - as is maintaining the general noisiness and 'messiness' of raw data. Modern privacy techniques can keep your data as close as possible to its raw format - minus the sensitive bits. That means preserving, correlations, aggregate values, and data distribution patterns without which your Test and Dev data would be nowhere near as rich and complex as the raw set.

4. It supports the test cases that are relevant for your production environment

All of the above combine to create data that replicates the instances that might cause issues in production. Unless your Test and Dev data resembles the nature and complexity of your raw data – and evolves with it, it won't allow you to realistically test for potential issues, such as the edge cases that representative testing needs (e.g. really long names or unusual payment types).

5. It's readily available whenever you need it

Traditionally, one of the biggest time sinks in getting data to Test and Dev was the manual process: a team would implement data privacy controls themselves or go through a compliance protocol – tying up resources and delaying speed-to-data. Modern privacy technology can get rid of such lengthy procedures: it can automate the application of policies.

This enables consistent privacy preservation at scale, for all new datasets. In short: with an automated, consistent provisioning pipeline, privacy-safe data gets to Test and Dev really quickly (e.g. in case of an emergency that needs an immediate fix).

6. It can safely be shared among teams and with external partners

The latest advances in privacy engineering include the concept of publishing de-identified data to a specific user group or for a specific purpose only (at Privitar, we call this a "Protected Data Domain", or PDD). The data inside a PDD can be joined and referential integrity will be preserved. However, data outside a PDD can't be linked to it. This limits use of the data to the specified purpose, massively reduces the risk of revealing sensitive information through linkage attacks, and allows organisations to safely share data with teams and partners (e.g. for integration testing or a PoC).

In Summary

There's no reason to use raw data at all anymore. Privacy engineering has come a long way. Modern data protection techniques are so sophisticated, they can deliver privacy-safe data for Test and Dev that preserves the characteristics of production data, while eliminating the enormous business risk of using raw data in such environments.

Best of both worlds.

We're Privitar

We help organisations engineer privacy-preserving data operations, using advanced data privacy techniques that protect sensitive information while retaining data utility. Our software accelerates and automates privacy-safe data provisioning, helping our customers get more business value from their data, generate data-driven insights, and drive innovation.

Contact us:

e: info@privitar.com
t: +44 203 282 7136
w: www.privitar.com

PRIVITAR

@PrivitarGlobal

www.privitar.com