clearbox AI
humans && machines

DATA CLONING FOR PRIVACY PRESERVATION

Synthetic data for better software testing

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai

Introduction

Adopting DevOps best practices is becoming paramount in big and small organisations to increase deployment frequency while reducing the number of issues showing up in production. **Continuous testing** is one of the essential elements to achieving that. Ideally, **companies should test software on real-life data**; however, it is difficult in many circumstances, especially when dealing with personal information data.

During this project an IT banking service provider, dealing with large amounts of personal data, used the Clearbox Synthetic Data Engine to build testing pipelines based on synthetic data.

CHALLENGE

Preparing data to feed test engines can take several weeks. This test data needs to be stored over time for non regression testing.

SOLUTION

The synthetic data engine was used by the client to create an ephemeral database the could be queried to generate test data on the fly.

RESULT

The client could obtain high quality test data in a matter of days instead of weeks. Furthermore by using the data engine they reduced test data storage costs by 90%.

Challenge

As software products evolve, testing them becomes more complex due to the growing number of components and microservices involved. It's essential to ensure that a product's behaviour remains consistent, for instance, after introducing a new user interface. Ideally, each software component, as well as the entire product, should be tested using real-life data to ensure accuracy and performance.

However, real-life data often contains sensitive personal information, making its use in testing limited by regulations like GDPR. For this reason, **data provisioning** for testing purposes is **a time-consuming process**, often delaying the testing workflow. Furthermore, non-regression testing, which verifies that existing functionalities are not broken by new changes, requires large sets of historical data. Storing these data sets long-term can be cumbersome and requires significant storage capacity, especially when maintaining comprehensive test environments for multiple software versions.

Solution

The organisation used our Synthetic Data Engine to ingest and clone their production database containing personal data. The cloning process took a couple of days and allowed the client to obtain a generator able to create synthetic data batches on demand. They used this generator as if it was a database containing on demand test data and were able to quickly populate a test engine. Additionally, since the generator creates data in a reproducible manner they decided to stop storing historical test data and to generate it on the fly using the generative model, providing both flexibility and significant storage savings while still maintaining test integrity.

Result

Creating realistic data for software testing allowed the organisation to improve their Continuous Integration/Continuous Delivery processes. The client could obtain high-quality test data in a matter of days instead of weeks. A virtually unlimited flow of realistic data enabled them to define a testbed for more granular tests while complying with data privacy regulations. Furthermore, by using the data engine instead of storing test data batteries, they reduced test data storage costs by 90%, thanks to the generative model's light storage requirements, which further optimised the testing process.

humans && machines

C.so Castelfidardo,30/a
10129, Torino (Italy)
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai