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

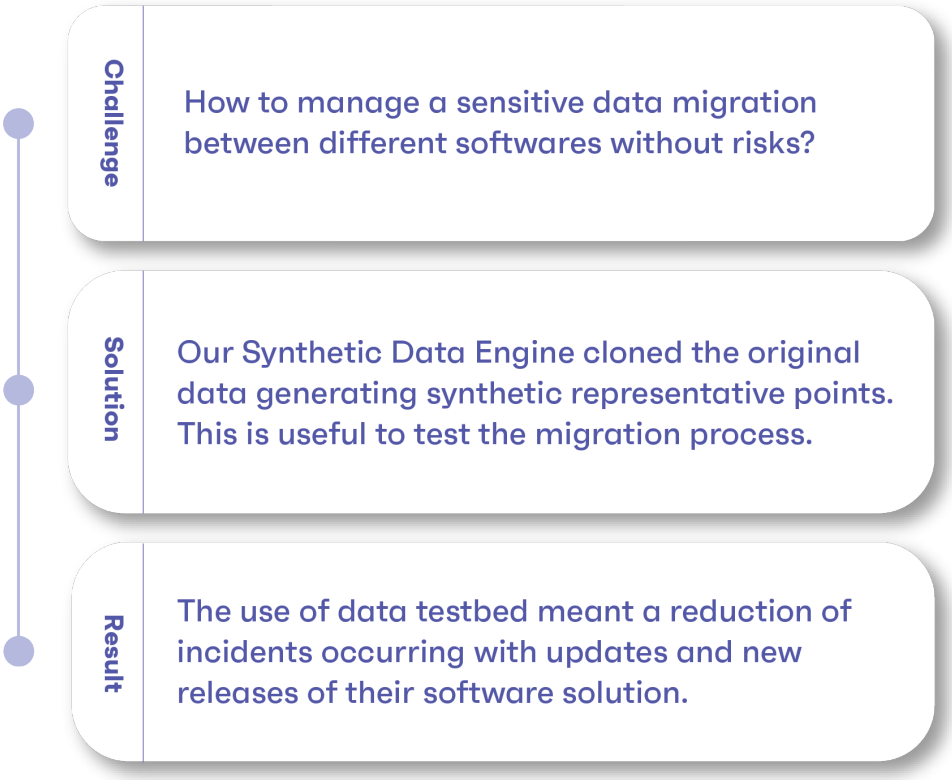
VAT ID: (IT)12161430017

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 use case, a public organisation dealing with large amounts of personal data used the Clearbox Synthetic Data Engine to build testing pipelines based on synthetic data.



Challenge

Testing software is becoming increasingly **complex** as the number of components and microservices used within IT products increases. We might be interested, for example, in checking that the behaviour of a product did not change after **migrating to a new infrastructure** or that a **new user interface is properly working** before it goes into deployment. Ideally, we should test each software component and the product as a whole **using real-life data**. Unfortunately, this is often impossible as real-life data usually contain personal information making its testing usage limited by regulations such as **GDPR**.

For this particular challenge, a governmental organisation had to migrate a database containing personal data to a new cloud provider. The operation presented a risk as many internal business processes were built on the database. The organisation wanted to make sure the operation would run smoothly by using **test data to populate the old and the new database** and to compare the behaviour of the old and the new system.

Solution

The organisation used our Synthetic Data Engine to **ingest and clone their production database containing individual data**. The cloning process generated several points representing **non-existing individuals** while **preserving the statistical properties** of the population from the original database. They finally injected the synthetic population both in the legacy and the new infrastructure databases and compared the behaviour of the two different software versions.

Result

Creating **realistic data for software testing** allowed the organisation to **improve their Continuous Integration/Continuous Delivery processes**. A virtually unlimited flow of realistic data allowed them to **define a testbed** for more granular tests while complying with data privacy regulations. The availability of such a data testbed corresponded to a **reduction of incidents** occurring with updates and new releases of their software solution.

humans && machines

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

VAT ID: (IT)12161430017

clearbox.ai