SYNTHESIZED | NextWave

Enable Best Data Practices with AI-powered Data Assets

Fast paced digital innovation is one of the key themes in Financial Services in 2020. COVID-19 has further driven the importance of digital and the need to innovate, faster and more effectively. However, many of the large financial institutions remain hamstrung when innovating. Their Achilles heel? Data.

Innovating In The Mud

Whilst new technologies offer immense promise of a better customer experience and a reduced operating cost, coupled with design thinking sessions that generate new product ideas and revenue streams, the innovation cycle and the initial promise comes to a grinding halt when large amounts of data are required to test and deploy new solutions. To date, there has only really been one answer to sourcing this data and therefore enabling development and innovation. **Existing client data**.

Client data is closely guarded in Financial Institutions. Access is limited and there are extensive controls and detailed risk acceptance procedures to follow just to get access to the data.
In some cases, organisations point blank refuse to allow production data to be deployed in test environments. All of this poses a number of challenges, as the development team quickly finds themselves in meeting after meeting seeking the approvals needed to use the data from departments ranging from Compliance, to Legal to Audit and Cyber Security.

The end result?

The extensive form filling activities mean organisations quickly lose momentum and fail to innovate quickly, resulting in lost business opportunities and a poorer customer experience.

All of this points to a process that is inefficient, lengthy, costly and does not facilitate nor encourage innovation at pace.

The Wider Picture

The approach to using client data is not just restricting innovation and resulting in loss of revenue. The reputational risk of using client data for testing remains at the front of mind for Chief Information Officers in financial institutions, being well aware of the serious risks and reputational damage that the organisation runs when a customer data breach occurs.

The current approach also constrains testing. Data cannot be shared with third parties, such as FinTechs, either due to legal constraints or the fear of data leakage. As such integration testing is restricted and the end results prone to errors that can impact the end user experience.

The fear of data leakage from test and development environments, and the requirement to hash key data fields results in unrepresentative data that can be used for model and system development testing. The effort required to produce the data set also means that it is used many times, and is often stale and outdated. As such testing phases and risk acceptance is completed on a subset of obscure data snapped at a specific time, meaning any recent changes in data which could impact the end solution are also ignored.

The move to the Cloud now means that client data often resides outside of the bank's premises. Whilst a number of banks continue to maintain ‘on prem’ data centres, the trend is to move to

external providers. Whilst organisations undertake extensive checks and implement controls, the fact remains that once client data leaves a production environment the risk of a data breach of the test data remains a real possibility.

With new rules and regulations governing security and privacy (GDPR) generating test data has become a real headache for financial institutions, that is absorbing valuable resources and time. GDPR cost companies not only more than £200 million in 2020 but also loss of reputation and brand damage.

Inherent Bias In Existing Client Data

Bias is inherent in customer data, and this bias can skew the results of testing and impact how the end product is designed and implemented. An organization's own data will have biases due to targeted customer groups, policy decisions and risk limits which become self-reinforcing and self-perpetuating when training models on existing data. Implicit bias also exists from the unknowns in the broader customer base which historically does not engage with the organisation. This makes the production based data sets ineffective, especially when the data is used to train machine learning models or to test algorithms, or when building and testing trading applications. Quite often the inherent bias and the impact of using the data is only identified late.

According to the recent independent report published by the Centre for Data Ethics and Innovation "We now have the opportunity to adopt a more rigorous and proactive approach to identifying and mitigating bias in key areas of life. Good use of data can enable organisations to shine a light on existing practices and identify what is driving bias. There is an ethical obligation to act wherever there is a risk that bias is causing harm and instead make fairer, better choices."

Lack of sufficient data / attributes

Volume is also a key factor, coupled with a requirement for key attributes when testing models and applications. Replicating data to generate volumes, again results in a non realistic, not to mention the time and effort required to generate the data set in the first place. In addition, you don't know what you don't know. Narrow data sets rarely mitigate and expose the edge case conditions that result in system failures and outages.

Considerate of these key challenges financial institutions face, we've looked for data visionaries that can provide viable solutions today. We partner up with Synthesized, a fast growing organization set to provide the world's fastest self service data mesh for data driven companies and this way, breaking silos, unlocking new streams of data products while guaranteeing compliance with the regulatory standards.

We believe synthetic data represents a revolution in testing and offers the potential to turbo change innovation. We have identified three key steps.

Step 1 — Free your data, and your staff

Synthesized data assets are totally new data sets that look, feel and behave like the original data set but are not replicating any of the detailed data itself. The data can be shown to have the same characteristics, signals, attributes and ranges without exposing the original data set beyond the synthesis point in an organisation.

Once created through the Synthesized engine, the data can be placed into user acceptance, test, development or sandbox environments with no risk of data leakage compromising the original customer or corporate data assets.

These new data assets address the problem of volume and variety for sparse, nonexistent or difficult to get data. Synthesising the data allows speculative uses to be investigated and validated before committing to the time consuming processes of data owner approvals, cross-border or cross-division approvals and legal approvals. The final testing will always need to be done on production data for final validation but only needs to be committed once the value case has been proven and potentially already iterated multiple times.

Sharing data with third parties to leverage Fintech advances, academic initiatives or evaluate new market advances become significantly lower risk as the data no longer contains traceable information or disclosure. Again, removing the cost, time and risk from the data sharing enables perspective and opportunistic innovation.

Step 2 — Understand and fix the weaknesses in your data

The benchmarking process will indicate any concentrations, biases, correlations and gaps in data providing scores and indicators. The Synthesized data assets can then be treated with this as a known, or rebalanced to eliminate or smooth out any of these features.

Step 3 — Scale up and stress your data

With the characteristics of these new data assets understood it is now possible to build data sets many magnitudes the size of the original test data. This allows for volume and scale testing of systems or models using representative data which may not have been encountered before but is possible. Stresses can be applied to the characteristics of the data to see what happens to models or systems when skewed data is fed in, simulate market stresses and extreme moves to see the impact on model performance with indicative data."

SYNTHESIZED | NextWave

What Happens Today

There are many other benefits of using Synthesized data. Examples include:

- **Protecting other types of sensitive data (as well as client data).** For example, transaction data when testing algo trading platforms or employee data if workflow led innovation is the focus.
- **Overcoming internal control barriers that block the innovation process.** These may be related to legal constraints and/or concerns over data leakage but it's the internal control framework that the control teams tend to focus on in discussions about providing access to data. Given that internal controls were originally designed around an on prem environment

they may be lagging, which can add complexity when teams want to utilise cloud solutions for innovation.

- **Additional volume related challenges.** Large volume test cases cause issues where financial institutions do not have spare compute capacity available on premises to run the scenarios they want to run. As such, they move innovation and testing to the cloud and by doing so increase the risk of data leakage. Innovation requiring low latency solutions (e.g. real time fraud checks) could be another example where the processing power is a key goal in the innovation, and this can only be achieved in the cloud.
- **Keeping it real.** Using unbiased, realistic data helps develop a better solution and end-user experience. This is particularly the case where they may have been working in companies with different businesses (it helps get over the *...it won't work for us because we are different...* discussion!).
- **Collaborate more effectively.** Synthetic data also offers a means of allowing financial institutions more options to collaborate with academia. Major banks and insurers now look to partner with Institutions to develop expertise but the biggest barrier to all of this is data sharing. Synthesized data allows research to be conducted on 'real' shapes of data — but without the issue of data leakage / privacy.

Is it even possible?

Yes, here's how:

The Synthesized Data Platform uses cutting-edge AI to automate all stages of data provisioning and data curation, synthesizing exactly the data required, thus enabling a real time data delivery process for development and testing in under 10 minutes without compromising on privacy or regulatory compliance.

Customers save hundreds of hours per project, see up to 80% in productivity gains, millions of pounds in data eng savings and increase in revenue.

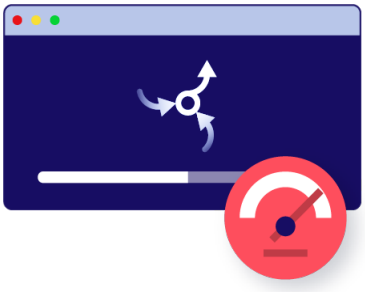
The Synthesized platform:

- Enhances, regularises and rebalances structured data automatically
- Enables more accurate models of risk to be developed
- Automates collaboration with partners on combating fraud and AML
- Accelerates all stages of joint projects with banks staying compliant with regulations
- Enables rapid testing against adversarial data scenarios
- Reduces the risk of mission-critical failure in production

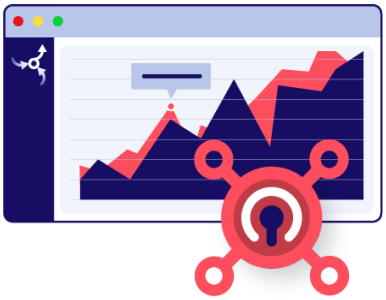
Leading financial institutions already use Synthesized for:

- Real time data curation, delivery and data value monitoring
- Collaboration with clients, partners and remote teams on data projects and pilots in a compliant manner
- Agile testing for functional and non functional in the cloud

- Keep employee data safe
- Keep transactional data safe
- Keep client data safe



- Blow away any cloud concerns
- Turbo charge innovation in the cloud
- Remove any data leakage worries
- Re-define the role of the CISO, CIO, CDO from contain and control to enable and deliver



- Collaborate with Academia and other large groups

Customer Story

A leading European retail bank slashed customer record access from months to hours with Synthesized while still remaining compliant.

The end result was faster data model development, wider access for collaboration, new external data sharing capabilities and generation of high quality test data.

The Project:

The bank wished to develop a model to determine the probability of certain customer behaviours in order to identify those with a high propensity to purchase and target them through a marketing campaign.

Time-to-insight was a major issue for the bank, with it taking an average of 2-6 weeks to get clearance to access customer data for development and analysis. In some cases, where specific banking regulations applied, it took many months to obtain access.

The objective was not only to reduce the time to access data from months to hours, but to broaden data access for more collaborative data projects, enhance public cloud sandboxes, share data safely externally and generate high quality test data.

Anonymization was another key concern; getting approval for data use is often conditional on individual identification being impossible, directly or indirectly. With concern about the security of traditional anonymization techniques on the rise, the bank needed to find a different solution.

The impact:

- Saved hundreds of hours per project
- Over 5 million rows of statistically representative data synthesized
- Models trained with data provisioned and preprocessed by Synthesized performed as well as, if not better than, those trained on original data
- Privacy by design — no occurrence of original sensitive data points

Moreover, Synthesized provided an always-on data project management dashboard connected to all data sources and allowing easy access to data projects. Projects were easily selected for movement across to data model deployment and visualization.

The bank carried out an evaluation of models trained using the Synthesized data versus the original data using statistical metrics. The models trained using synthetic data performed as well as, and in some cases better than, those trained on the original data.

SYNTHESIZED | NextWave

> “Synthesized is a leader in the field of high quality data provisioning. The results were superior to other solutions tested, executed extremely fast and saved hundreds of hours on one project alone” said the bank project champion.

Want to know more? Contact us at www.nxwave.com and www.synthesized.io