Towards open and agile Big Data analytics in Financial Sector

CaixaBank's Success Story in I-BiDaaS

Organizations leverage data pools to drive value, and it is variety, not volume or velocity, which drives big-data investments. This trend also applies to the banking sector, and CaixaBank has been developing its own big data infrastructure over many years and receiving several awards (e.g. “2016 Best Digital Retail Bank in Spain and Western Europe” by Global Finance) as a consequence. Indeed, CaixaBank is the third largest financial institution in Spain and is currently the leading force in Spanish retail banking, with almost 14 million customers across Spain and Portugal under their subsidiary brand BPI.

To offer high-quality services to its customers, ***CaixaBank has a network of more than 5,000 branches with over 40,000 employees and manages an infrastructure with more than 9,500 ATMs, 13,000 servers, and 30,000 handheld devices. This infrastructure results in a massive amount of data collected every day by all the bank systems and channels, aggregating information of CaixaBank operation from the clients, employees, third-party providers, and autonomous machines***. In total, CaixaBank has more than 300 different data sources used by its consolidated big data models and more than 700 internal and external active users enriching its data every day, which is translated into a Data Warehouse with more than 4 petabyte that increases 1 petabyte per year. ***Much of this information is already utilized by means of big data analytics techniques***, for example, to generate security alerts and prevent potential frauds -CaixaBank receives around 2,000 attacks per month- and did so before joining I-BiDaaS.

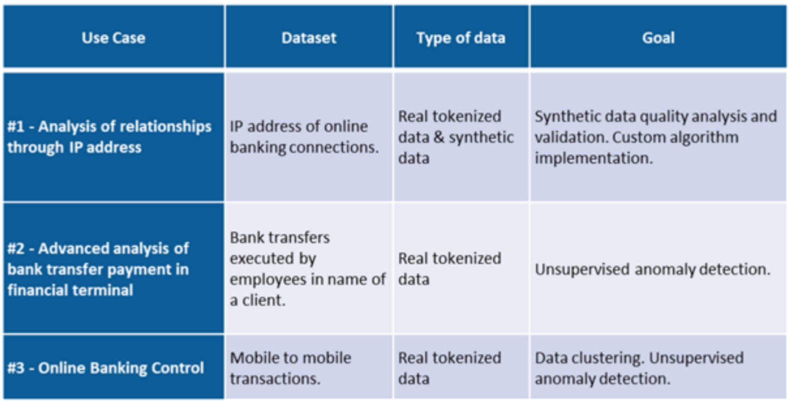
Why CaixaBank participation in I-BiDaaS was important and why it is resulting on a very successful story?

I-BiDaaS is a H2020 Research and Innovation project that aims to empower users to easily utilize and interact with big data technologies, by designing, building, and demonstrating, a unified solution that significantly increases the speed of data analysis while coping with the rate of data asset growth, and facilitates cross-domain data-flow towards a thriving data-driven EU economy. The project focuses on providing a self-service solution that ***will empower CaixaBank employees, who are the true decision-makers, giving them the insights and the tools they need to make the right decisions in a much more agile way.***

In the case of CaixaBank, as with many entities in critical sectors, there was initial reluctance to use any big data storage or tool outside its premises. Therefore, the ***primary goal of CaixaBank when starting its involvement in I-BiDaaS was to find an efficient way to perform big data analytics outside its premises***. Achiving this would speed up the process of granting new external providers access CaixaBank data (typically a bureaucratic process that takes weeks). Additionally, CaixaBank wanted to become ***much more flexible in adopting proof-of-concept (PoC) technological solutions (i.e., to test the performance of new data analytics technologies to be integrated into CaixaBank infrastructure)***. Usually, for any new technology testing, even simple ones, if hardware is needed, then it should be done through the infrastructure management subsidiary who will be in charge of deploying it. Due to the level of complexity, the size of CaixaBank’s infrastructure, and the processes rigidity, deployment can also take months.

CaixaBank needed to find ways to by-pass these processes without compromising security or privacy. GDPR really limits the usage of customer data, even if used for fraud detection and prevention or for enhancing the security of customer accounts. It can be used internally to apply certain security policies, but sharing this data with other stakeholders remains an issue. Furthermore, the banking sector is strictly regulated, and National and European regulators are supervising all security measures taken by banks to provide a good level of security while, maintaining the privacy of customers. The current trend of externalizing many services to the cloud also implies the establishment of strict control of the location of data as well as who has access to it.

The I-BiDaaS CaixaBank-roadmap (see Figure 1) had a turning-point, in which ***CaixaBank completely changed its approach approach from a non-sharing real data at all position to looking for the best way possible to share real data and perform big data analytics outside its facilities***. I-BiDaaS helped to push for internal changes in policies and processes and evaluate tokenization processes as an enterprise standard to extract data outside their premises, breaking both internal and external data silos.



Use Case	Dataset	Type of data	Goal
#1 - Analysis of relationships through IP address	IP address of online banking connections.	Real tokenized data & synthetic data	Synthetic data quality analysis and validation. Custom algorithm implementation.
#2 - Advanced analysis of bank transfer payment in financial terminal	Bank transfers executed by employees in name of a client.	Real tokenized data	Unsupervised anomaly detection.
#3 - Online Banking Control	Mobile to mobile transactions.	Real tokenized data	Data clustering. Unsupervised anomaly detection.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 780787

This project is part of BIG DATA VALUE PUBLIC-PRIVATE PARTNERSHIP

Results obtained from the first use case validated the usage of rule-based synthetically generated data and indicated that it can be very useful in accelerating the onboarding process of new data analytics providers (consultancy companies and tools). ***CaixaBank validated that it could be used as high-quality testing data outside CaixaBank premises for testing new technologies and PoC developments, streamlining the grant accesses of new external providers to these developments, and thus reducing the time of accessing data from an average of 6 days to 1.5 days***. This analysis was beneficial for CaixaBank purposes, ***but was also concluded that the analysis of rule-based fabricated data did not enable the extraction of new insights from the generated dataset***, simply the models and rules used to generate the data.

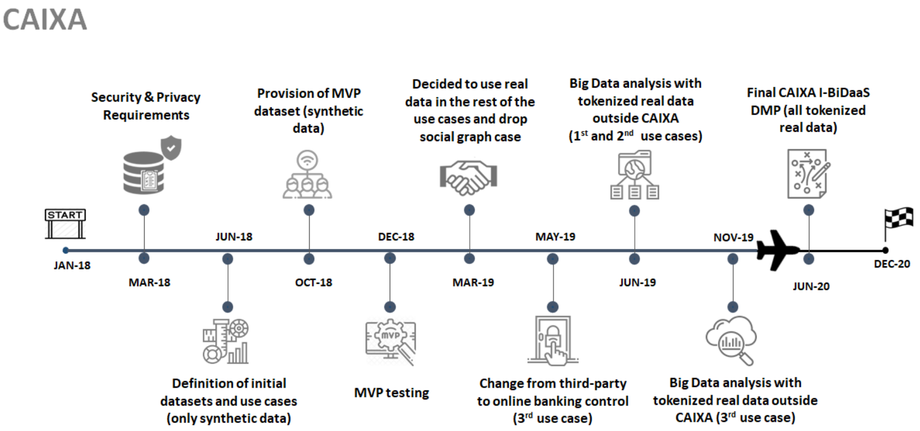

CAIXA
START
JAN-18
Security & Privacy Requirements
MAR-18
JUN-18
Definition of initial datasets and use cases (only synthetic data)
Provision of MVP dataset (synthetic data)
OCT-18
DEC-18
MVP testing
Decided to use real data in the rest of the use cases and drop social graph case
MAR-19
MAY-19
Change from third-party to online banking control (3rd use case)
Big Data analysis with tokenized real data outside CAIXA (1st and 2nd use cases)
JUN-19
NOV-19
Big Data analysis with tokenized real data outside CAIXA (3rd use case)
Final CAIXA I-BiDaaS DMP (all tokenized real data)
JUN-20
DEC-20

Figure 1. CaixaBank's roadmap in the I-BiDaaS project

The I-BiDaaS CaixaBank-roadmap (see Figure 1) was a turning point for ***CaixaBank, which completely changed its approach from a non-sharing real data at all position to looking for the best way possible to share real data and perform big data analytics outside its facilities***. I-BiDaaS helped to push for internal changes in policies and processes and evaluate tokenization processes as an enterprise standard to extract data outside their premises, breaking both internal and external data silos.

The other two use cases focused on how extremely sensitive data can be tokenized to extract real data for use outside CaixaBank premises. By tokenizing, we mean encrypting the data and keeping the encryption keys in a secure data store that will always reside in CaixaBank facilities. This approach implies that the data analysis will always be done with the encrypted data, and it can still limit the results of the analysis. One of the challenges of this approach is to ***find ways to encrypt the data in a way that it loses as little relevant information as possible***. Use case 2 and use case 3 experimentation was performed with tokenized datasets built by means of three different data encryption algorithms: (1) Format preserving encryption for categorical fields; (2) Order preserving encryption for numerical fields; (3) A Bloom-filtering encryption process for free text fields. This enabled CaixaBank to extract the dataset, ***upload it to I-BiDaaS self-service big data analytics platform and analyse it with the help of external entities without being limited to the corporate tools available inside CaixaBank facilities.*** I-BiDaaS Beneficiaries proceeded with an unsupervised anomaly detection in those use cases, identifying a set of pattern anomalies that were further checked by CaixaBank's Security Operation Center (SOC). This helped increase the level of financial security of CaixaBank. However, beyond that, we consider this experimentation very beneficial, and should be replicated in other commercial big data analytics tools, previously to their acquisition. In summary, the next table highlights some of the benefits of CaixaBank due to its participation in I-BiDaaS:

Benefits	KPIs
To increase the efficiency and competitiveness in the management of its vast and complex amounts of data.	75% time reduction data access from external stakeholders using synthetic data (From 6 to 1.5 days).
To break data silos not only internally, but also fostering and triggering internal procedures to open data to external stakeholders.	Real data accessed by at least 6 different external entities skipping long-time data access procedures.
To evaluate Big Data analytics tools with real-life use cases of CaixaBank in a much more agile way.	I-BiDaaS overall solution and tools experimentation with 3 different industrial use cases with real data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780787

This project is part of BIG DATA VALUE PUBLIC-PRIVATE PARTNERSHIP