Data Sharing and Data Availability: Lessons Learned

Organizations worldwide seek knowledge in order to develop competitive advantage in the Information Age and the convergence of IoT, cloud, and big data offers opportunities to this end. However, the exploitation of big data technologies is, at times, extremely expensive in terms of funding, assets, and workforce.

On such basis, I-BiDaaS aims to facilitate utilization of big data technologies by providing to organizations a self-service solution that will empower their employees with the right knowledge and give the true decision-makers the insights they need to make the right decisions. Profitability, cost reduction and employees' empowerment are some of the impacts that I-BiDaaS will bring to the organizations, providing them eventually with the competitive advantage they need towards a thriving data-driven EU economy.

In view of the plethora of domains that can exploit such self-service solutions, I-BiDaaS explores three critical ones with significant challenges and requirements: banking, manufacturing, and telecommunications. After a significant period of research, design and experimentation towards the development of the I-BiDaaS solution, the project's data providers are now able to share their experience regarding the challenges, difficulties and obstacles they faced on data availability and data sharing.

Their stories have been transformed to lessons that provide valuable knowledge to our stakeholders and also necessary insight on how to further improve the I-BiDaaS solution.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

CaixaBank: I-BiDaaS Application to the Financial Sector

In the case of CaixaBank, as with many entities in critical sectors, there was an 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**. Achieving this would speed up the process of granting new external providers access to 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.

To be more efficient in terms of amount of money and time spent, CaixaBank proposed three different use cases to be tested within I-BiDaaS, exploring ways to conduct data analytics outside of its premises, assuring the maximum level of security and privacy possible and evaluating I-BiDaaS overall solution and specific tools, such as IBM's TDF (Test Data Fabrication) for providing synthetic, non-sensitive data.

Results obtained from the first use case validated the usage of rule-based synthetically generated data and indicated that it could 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.

The other two use cases focused on how extremely sensitive data can be tokenized to extract real data for use outside CaixaBank premises. By

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

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**. We proceeded with 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 provide us the guidelines to be replicated when analysing 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.

TELEFONICA I+D: I-BiDaaS Application to the Telecommunication Sector

There are **several challenges** related to Telefonica's setup. One of the most critical ones is the **use of big data outside the company's premises**. The I-BiDaaS initiative served as a **pivoting point** for re-visiting existing practices of performing **big data analytics within a wider ecosystem**.

More specifically, when considering the telecommunications sector, the utility of data generated across the network depends on how the data are aggregated across time, users, locations involved, etc. Depending on where and how the aggregation and modeling of such data happens, it can be extremely powerful for marketing and other purposes (e.g., infrastructure planning) but also raise great privacy concerns over the anonymity and privacy of the users producing them. For example, on the one hand, browsing data of particular users can be very valuable for targeted advertisements but also be considered as very invasive and creepy by some users. On the other hand, aggregated mobility data of users in a city extracted from antenna logs can be useful in deploying new infrastructure, services, bootstrapping other telco companies, or even planning physical retail stores without violating users' privacy due to aggregation.

Second, leakage of private data happening over the network can impact or diminish the utility of such data, and threaten the users involved, as well as their trust in the ISP and its services. In addition, sharing across companies of insights or data based on user information and activity is regulated and can be difficult to achieve, again diminishing the utility of such data and insights. Also, such sharing is not easy to audit and monitor, especially when this needs to be done at the user-level, or at least with the user's consent. This is especially true if these data are user-generated (e.g., web browsing logs), and must be stored, processed and shared in a GDPR-compliant manner.

In fact, insights based on machine learning models are typically very useful for telcos and other organizations (for the reasons mentioned above), but again **very difficult to audit and even engage the users to participate** and control the model creation. In particular, it is difficult to build a machine learning model across many users' data while respecting their privacy, possible opt-ins in particular analysis to be performed on their data, etc.

These challenges become even more difficult to address when data are combined from different sources within a company, or taken to the extreme case, combined across companies. Therefore, in order for such diverse valued data to be utilized in a **secure, privacy-preserving, industry-supported, future EU data market, safeguarding policies and procedures are put in place**, and innovative technologies are needed in the near future. In more detail, privacy-preserving machine learning methods applied to the data beforehand could enable higher or even exact compliance with GDPR-type of regulations, and provide better privacy guarantees to the end-user. Also, user models built with a privacy-preserving fashion could be traded or shared with 3rd party companies (e.g., other telcos, advertisers, etc.) **via the 4th platform strategic effort of TID**.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

In many cases, an anonymised dataset can still present residual risk to data subjects. Indeed, even when it is no longer possible to precisely retrieve the record of an individual, it may remain possible to glean information about that individual with the help of other sources of information that are available (publicly or not).

It has to be highlighted that beyond the direct impact on data subjects produced by the consequences of a poor anonymisation process (annoyance, time consumption and feeling of lost control by being included in a cluster without awareness or prior consent), other indirect consequences of poor anonymisation may occur whenever a data subject is included in a target erroneously by some attacker, as a consequence of processing anonymised data - especially if the attacker's intents are malicious. Therefore, anonymisation techniques can provide privacy guarantees, but only if their application is engineered appropriately - which means that the prerequisites (context) and the objective(s) of the anonymisation process must be clearly set out in order to achieve the targeted anonymisation level.

Users interact with various telco services and trigger the production of large and diverse types of data at different levels in the network infrastructure of the ISP. Some data are produced at the user level (i.e., end-user devices such as smartphones, laptops, etc.), and include data regarding activities of the user such as web browsing, TV viewing, radio streaming, video calling, etc. Other data are produced at the networking infrastructure level (e.g., routers, antennas, etc.), as a reaction to the user activity in the network such as phone calls, activity with SMS and other message services, roaming, traveling in the city (i.e., changing antennas), web browsing, and other user-level activities.

Both types of data are useful for modeling user behavior, and can be utilized by the telco owning them, or by other telcos and 3rd party companies. However, with **strict EU regulations on user personal data such as the GDPR and ePrivacy**, safeguarding methods were needed to put in place in order to prevent privacy-breaching activities. To this end, TID has made **concrete steps to utilize its data by providing new services on top of them, or supporting other 3rd party businesses in making efficient and effective, data-driven decisions**.

Towards this goal, **TID created a novel strategy called the 4th platform**, and a first of its kind vertical on top of it, called Aura, which aims to revolutionize the way telcos and other big data players can participate in the data markets in a user privacy-respecting fashion. Moreover, in the context of I-BiDaaS, TID has provided the necessary support so that IBM can deliver synthesized data **to support the relevant use cases at early-stage**. In addition, TID has **achieved the data silo breach and has successfully shared resources with authorized third parties**.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

CRF: I-BiDaaS Application to the Manufacturing Sector

Automotive production processes are complex in that production lines have several robots and digital tools. At the shop floor level, massive amounts of raw data are gathered; data that not only help to monitor processes but can also improve process robustness and efficiency.

Within the **I-BiDaaS project**, the data provider CRF identified two scenarios in which complex and initial structured/unstructured data sets are retrieved from real processes, and defined two use cases:

1. Production process of Aluminium die-Casting
2. Maintenance and Monitoring of Production assets

The project focuses on providing **a self-service solution** that will give CRF employees the insights and tools they need **to develop a methodology to be implemented in the production sites for improving the quality of the processes and products in a much more agile way** through the collaborative effort of self-organizing and cross-functional teams.

1. Production process of Aluminium die Casting

The Aluminium die-casting process is complex, and it is important to not only carefully design parameters but also to control them because they have a direct impact on the quality of the casting. The goal of the use case is **to predict whether an engine block will be manufactured correctly**, i.e., without anomalies. The correct prediction during the die-casting process would **avoid subsequent further processing and scraps**, **which would lead to financial savings for the manufacturer.**

Initially, we received large unstructured volumes of heterogeneous data from different sources and at different levels. In the preliminary stage, we analysed all the information and interacted with the plant in order to understand the obscure parts. Furthermore, needing to guarantee the anonymization of the data, a process that took more time than expected, synthetic data were fabricated in order not to delay the technical development of the individual components and the I-BiDaaS solution. Later, when real anonymised data became available, we shared them to test the complexity of the process with the analytics developed within the project.

Unstructured, noisy and incomplete data has been collected, aggregated, pre-processed and converted into structured messages of a common, unified format for the analysis. In summary, the first lesson we learned is that, in parallel with the process of data anonymization, creating structured data, etc., it is useful to generate synthetic data in parallel. This is useful for the early stages of development but requires caution when extracting insights from synthetic data. **The high-level algorithms**, developed during the first part of the project, **identified the critical values from the dataset and determined the parameters that affect the quality of the process**. According to the first results, we proposed to the plant some strategies for better control of data integrity (e.g., a second level of control for changes in the parameters of the operators) and two thermal imaging cameras have

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

been installed in all die-casting machines. Therefore, a large dataset of thermal images has been provided in addition to the process data, assuming that there is a correlation between the sensor data, the thermal data and the process result. Several models have been developed to utilize both sensor and thermal image data, as reported in D3.3, available on the I-BiDaaS' website. The second lesson we learned is that the initial data analysis, jointly with the CRF hackathon results (see the related blog article 'I-BiDaaS – CRF Hackathon'), revealed that an additional dataset of different nature is to be included due to inherent performance limitations. This was not possible to assess beforehand, and the benchmarking and analysis through the hackathon helped to improve the process. In summary, it may be worthwhile for the data collection process-experiment design process and the data analysis process be iterative rather than sequential "in one shot".

The results can be made available to the end-user via the AEGIS's Advanced Visualization module on which the I-BiDaaS User Interface is built upon and allows us to timely check the status of the process and to classify the quality levels that we are using as KPIs. The third lesson we learned is that Advanced visualizations are extremely useful for developing high value Big Data analytics solutions for domain experts and operators.

2. Maintenance and Monitoring of Production assets

The Maintenance and monitoring of production assets use case gathers data from the production process to define timely "predictive" maintenance measures **to identify and prevent failures before they occur. Different sensors are installed on the production line**, acquiring different types of data (e.g., sensors mounted on different machines, accelerometers mounted in linear stages). The sensors are connected to the control units on the line, while the control units are connected to a data server. The data consists of two different datasets, the SCADA and the MES. The SCADA dataset contains production, process and control parameters of the daily vehicle production. The MES data contains specific data associated with the type of vehicle being produced.

Initially, our intention was to use both types of data, but over time, we faced **problems retrieving MES data because of the scheduled activities and changes in the production lines, partially due to the Covid-19 Pandemic**. Therefore, we decided to utilize only SCADA data to obtain thresholds for abnormal measurements for all sensors. The fourth lesson we learned is that when defining an industrial use case, it turns out that it is very important to allow flexibility and redundancy in the use case definition so that the work can proceed under unexpected or severe conditions, including subsystem failure, unavailability of data, etc.

Due to our internal constraints, the lack of the necessary infrastructure and the high volume of available data, the **I-BiDaaS technical partners decided to bridge the I-BiDaaS infrastructure with the CRF internal server for real-time data transfer and near to real-time data analysis.**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

Find & Follow us

Website | Twitter | LinkedIn

Zenodo | OpenAIRE | GitHub

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.