

Chapter

Big Data Analytics in the Banking Sector: Guidelines and Lessons Learned from the CaixaBank Case

Andreas Alexopoulos¹, Yolanda Becerra², Omer Boehm³, George Bravos⁴, Vasilis Chatzigiannakis⁴, Cesare Cugnasco², Giorgos Demetriou⁵, Iliada Eleftheriou⁶, Lidija Fodor⁷, Spiros Fotis¹, Sotiris Ioannidis^{8,9}, Dusan Jakovetic⁷, Leonidas Kallipolitis¹, Vlatka Katusic⁵, Evangelia Kavakli⁶, Despina Kopanaki⁹, Christoforos Leventis⁹, Mario Maawad Marcos¹⁰, Ramon Martin de Pozuelo¹⁰, Miquel Martínez², Nemanja Milosevic⁷, Enric Pere Pages Montanera¹¹, Gerald Ristow¹², Hernan Ruiz-Ocampo⁵, Rizos Sakellariou⁶, Raül Sirvent², Srdjan Skrbic⁷, Ilias Spais¹, Giorgos Vasiliadis⁹, Michael Vinov³

¹Aegis IT Research LTD, UK ²Barcelona Supercomputing Center, Spain ³IBM, Israel

⁴Information Technology for Market Leadership, Greece ⁵Ecole des Ponts ParisTech, France

⁶University of Manchester, UK ⁷University of Novi Sad - Faculty of Sciences, Serbia ⁸Technical

University of Crete - School of Electrical and Computer Engineering, Greece ⁹Foundation for

Research and Technology, Hellas - Institute of Computer Science, Greece ¹⁰CaixaBank, Spain

¹¹ATOS, Spain ¹²Software AG, Germany

Abstract. A large number of EU organizations already leverage Big Data pools to drive value and investments. This trend also applies to the banking sector. As a specific example, CaixaBank currently manages more than 300 different data sources (more than 4 PetaBytes of data and increasing) and more than 700 internal and external active users and services are processing them every day. In order to harness value from such high-volume and high-variety of data, banks need to resolve several challenges, such as finding efficient ways to perform Big Data analytics and to provide solutions that help to increase the involvement of bank employees, the true decision makers. In this book chapter, we describe how these challenges are resolved by the self-service solution developed within the I-BiDaaS project. In more detail, we present three CaixaBank use cases, namely i) *Analysis of relationships through IP addresses*; ii) *Advanced Analysis of bank transfer payment in financial terminal*; and iii) *Enhanced control of customers to online banking*, and describe how the corresponding requirements are mapped to specific technical and business KPIs. For each use case, we present the architecture, data analysis and visualization provided by the I-BiDaaS solution, reporting on the achieved results, domain-specific impact, and lessons learned.

Keywords: Self-service solution, Banking, Security applications, Big data analytics, Advanced analytics and visualizations

1 Introduction

Collection, analysis and monetisation of Big Data is rapidly changing the financial services industry, upending the longstanding business practices of traditional financial institutions. By leveraging vast data repositories, companies can make better investment decisions, reach new customers, improve institutional risk control and capitalise on trends before their competitors. But given the sensitivity of financial information, Big Data also spawns a variety of legal and other challenges for financial services companies¹.

Following this digitalisation trend, CaixaBank has been developing its own Big Data infrastructure since years ago, being awarded several times (e.g. “2016 Best Digital Retail Bank in Spain and Western Europe” by Global Finance). With almost 14 million clients over Spain (and Portugal under their subsidiary brand BPI), 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 handhelds. All those figures represent a massive amount of data collected every day by all the bank systems and channels, gathering relevant information of the bank operation from the clients, employees, third-party providers and autonomous machines. In total, CaixaBank has more than 300 different data sources used by their consolidated Big Data models and more than 700 internal and external active users enriching their data every day, which is translated into a Data Warehouse with more than 4 PetaBytes (PBs), which increases 1PB per year.

Much of this information is already used in CaixaBank 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-. Anyway, CaixaBank is one of the banking leaders in the European and national collaborative research, taking part in pre-competitive research projects. Within the EU I-BiDaaS project (funded by the Horizon 2020 Programme under Grant Agreement 780787), CaixaBank identified three concrete use cases, namely *i) Analysis of relationships through IP addresses; ii) Advanced Analysis of bank transfer payment in financial terminal; and iii) Enhanced control of customers to online banking* to study the potential of a Big Data self-service solution that will empower its 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 rest of this chapter, Section 2 discusses the requirements and challenges for Big Data in the banking sector. Section 3 details the different use cases considered, together with their technical and business KPIs. In Section 4, for each use case, we present the architecture, data analysis and visualisation of the I-BiDaaS solution, reporting on the achieved results and domain-specific impact. It also relates the described solutions with the BDV reference model and priorities of the BDV Strategic and Research Innovation Agenda (SRIA) [1]. Section 5 summarises the lessons learned through all the experiments deployed by CaixaBank and the rest of I-BiDaaS partners, and finally, Section 6 presents some conclusions.

¹ https://www.wilmerhale.com/uploadedFiles/Shared_Content/PDFs/Services/WilmerHale-Big-Data-FinancialServices.pdf

2 Challenges and Requirements for Big Data in the banking sector

The vast majority of banking and financial firms globally believe that the use of insight and analytics creates a competitive advantage. The industry also realises that it is sitting on a vast reservoir of data and insights can be leveraged for product development, personalised marketing, and advisory benefits. Moreover, regulatory reforms are mainly leading to this change. Ailing business and customer settlements, continuous economic crisis in other industry verticals, high cost of new technology and business models, and high degree of industry consolidation and automation are some of the other growth drivers. Many financial services currently focus on improving their traditional data infrastructure as they have been addressing issues such as customer data management, risk, workforce mobility, and multichannel effectiveness. These daily problems led the financial organisation to deploy Big Data as a long-term strategy and it has turned out to be the fastest growing technology adopted by financial institutions over the past five years².

Focusing on the customer is increasingly important and the critical path towards this direction is to move the data analytics tools and services down to the employees with direct interaction with the customers, utilising Big-Data-as-a Self-Service solutions³ [2].

Another critical requirement for financial organisations is to use data and advanced analytics for fraud and risk mitigation and achieving regulatory and compliance objectives. With cyber security more important than ever, falling behind in the use of data for security purposes is not an option. Real-time view and analysis is critical towards competitive advantage in the financial/banking sector.

The usage of Big Data analytics is gradually being integrated in many departments of the CaixaBank (security, risks, innovation, etc.). Therefore, there is a heterogeneous group of experts with different skills but also relies on several Big Data analytics experts that provide consultancy services. However, the people working with the great amount of data collected from the different sources and channels of CaixaBank can be grouped into the following categories (which indeed could be fairly generalised to other financial entities):

- IT & Big Data expert users: employees and third-party consultants with excellent programming skills & Big Data analytics knowledge.
- Intermediate users: People with some notion on data analytics that are used to work with some Big Data tools, especially for visualisation and Big Data visual analysis (such as QlikSense/QlikView⁴). They are not skilled programmers, although they are capable of programming simple algorithms or functions with Python or R.
- Non-IT users: People with an excellent knowledge of the field and the sector; they could interpret the data, but they lack programming skills or Big Data analytics knowledge.

² <https://www.mordorintelligence.com/industry-reports/big-data-in-banking-industry>

³ <http://www.gartner.com/it-glossary/self-service-analytics>.

⁴ <https://www.qlik.com/>.

Although ‘IT & Big Data expert users’ are getting more involved and being a relevant part of the business day-by-day operations of the entity, there are few compared to the ‘Intermediate’ and ‘Non-IT’ users’. Reducing the barriers and the knowledge required by those user categories in exploiting efficiently the collected data represents one of the most relevant challenges for CaixaBank.

With all the above mentioned, the I-BiDaaS methodology for eliciting CaixaBank requirements (see Table 1) took into consideration the specific challenges faced by CaixaBank, as well as the literature on Requirements Engineering (RE) approaches specifically for Big Data applications [3].

Table 1. CaixaBank consolidated requirements

	<i>Business Requirements</i>
R1	To speed up the implementation of new Big Data analytics applications (business goal).
R2	To be able to test new data analytics tools and algorithms outside CaixaBank premises whilst assuring maximum level of security/privacy (business goal).
R3	To enable third parties to efficiently implement and test new tools and algorithms without accessing real data (business goal).
R4	To ensure accuracy and reliability of analytics process (quality business goal).
R5	To improve efficiency of the analytics process (quality business goal).
R6	Time efficiency (KPI).
R7	Cost reduction (KPI).
	<i>User Requirements</i>
R8	Data is collected by several different sources (ATMs, online banking services, employees’ workstations, external providers’ activity, network devices, etc.) (data provider requirement).
R9	Data are owned by CaixaBank and are not publicly available. They can be shared with third parties only once the data is anonymised (data provider requirement).
R10	Support the use of techniques related to log analysis such as process mining algorithms or similar (Big Data analytics provider requirement).
R11	Users will be able to download results (in several formats such as .csv, .xls, etc.) in order to analyse them by their own or send them to other employees of the Security Operation Centre (data consumer requirement).
R12	Intermediate users will be able to modify parameters of the algorithms and refine the initial results (data consumer requirement).
	<i>System Requirements</i>
R13	The system should enable the generation of anonymised and synthetic data to enable safe experimentation and testing (functional requirement).
R14	The system should support diversified, analytic processing, machine learning and decision support techniques to support multiple stages of analysis (functional requirement).
R15	The system should ensure security of sensitive data (non-functional requirement).

In particular, the I-BiDaaS methodology followed a goal-oriented approach to requirements engineering [4] whereby elicitation of requirements was seen as the

systematic transformation of high-level business goals that reflect the company vision with respect to the Big Data analytics activity or project, to the user requirements of the groups of stakeholders involved (e.g., data providers, Big Data capability providers, data consumers) and finally to the specific system functional and non-functional requirements, which describe the behaviour that a Big Data system (or a system component) should expose, or the capabilities it should own in order to realise the intentions of its users.

The requirements elicitation process was carried out in collaboration with both CaixaBank stakeholders and Big Data technology providers. It involved two steps: the first step was to extract specific requirements based on the characteristics of each CaixaBank use case; the second step involved the consolidation of all requirements in a comprehensive list. Appropriate questionnaires were used to assist participants express their requirements. Requirements consolidation was guided by generic requirements categories identified through the review of RE works for big data applications [5].

Although described in a linear fashion, the above activities were carried out in an iterative manner resulting in a stepwise refinement of the results being produced. The complete list of all requirements elicited is described in detail in [6].

3 Use cases description and experiments' definition; technical and business KPIs

The CaixaBank experiments aim at evaluating and validating the self-service Big Data platform [9] proposed in the framework of the I-BiDaaS project, and its implementation in the specific CaixaBank use cases. More precisely, the experiments aim to test the efficiency of the I-BiDaaS platform for reducing the costs and the time of analysing large datasets whilst preserving data privacy and security.

The definition of the experiments follows a goal-oriented approach, whereby for each experiment: the experiment's goal(s) towards which the measurement will be performed are first defined; then a number of questions are formed aiming to characterise the achievement of each goal; and finally, a set of Key Performance Indicators (KPIs) and the related metrics are associated with every question in order to answer it in a measurable way.

Such KPIs have been defined at the business level during the user requirements elicitation phase (see Section 2). However, they need to be further elaborated and refined so that they can be mapped onto specific indicators at the Big Data application and platform level. This ensures that (a) both business and technical requirements are taken into consideration and (b) the traceability among business and application performance are taken into consideration. In addition, for each KPI, the baseline (current) value and the desired improvement should also be defined, whose measurement relates to the achievement (or not) of the specific indicator.

The definition of each experiment also included the definition of the experiment's workflow in terms of the type and order of activities (workflow) involved in each experiment, as well as the definition of the experimental subjects that will be involved in the experiment.

Taking all the aforementioned into account, CaixaBank proposed three different use cases and evaluated the I-BiDaaS tools from the perspective of potential usage by those different groups of employees:

- Enhance control of customers to online banking.
- Advanced Analysis of bank transfer payment in financial terminal.
- Analysis of relationships through IP addresses.

The rest of the section includes the final use cases definitions in chronological order as developed and deployed in the project. We also refer the reader to Section 4 for further details of the use cases corresponding solutions and Section 5, which provides a complementary description, collecting the lessons learned acquired during the respective processes.

3.1 Analysis of relationships through IP addresses

‘*Analysis of relationships through IP addresses*’ was the first use case selected to test the I-BiDaaS Minimum Viable Product (MVP). In this use case, CaixaBank aims to validate the usage of synthetic data and the usage of external Big Data analytics platforms. It is deployed in the context of identifying relationships between customers that use the same IP address when connecting to online banking. CaixaBank stores information about their customers and the operations they perform (bank transfer, check their accounts, etc.) using channels such as mobile apps or online banking, and they afterwards use this data for security and fraud prevention processes. One of the processes is to identify relationships between customers and use them to verify posterior bank transfers between linked customers. Such operations are considered with lower possibility to be fraudulent transactions. It allows CaixaBank’s Security Operation Centre (SOC) to directly discard those bank transfers during the revision processes. The goal of this experiment is to validate the use of synthetic data for analysis (i.e. one of the I-BiDaaS platform features), evaluate the quality of the synthetic data (i.e. if the algorithm can find the same amount and patterns of connections using the real and the synthetic datasets) and to test the time efficiency of the I-BiDaaS solution.

3.2 Advanced Analysis of bank transfer payment in financial terminal

The second CaixaBank use case that was studied in I-BiDaaS is “*Advanced Analysis of bank transfer payment in financial terminal*”. This use case aims to detect the differences between reliable transfers and possible fraudulent cases. The goal of this experiment is to test the efficiency of the I-BiDaaS solution in the context of anomaly detection in bank transfers from employees’ workstations (*financial terminal*).

For that reason, the first step was to identify all the contextual information from the bank transfer (i.e. time execution, transferred amount, etc.), the sender and receiver (e.g. name, surname, nationality, physical address, etc.), the employee (i.e. employee id, authorisation level, etc.) and the bank office (e.g. office id, type of bank office, etc.). All this information is coming from several relational database tables stored in the CaixaBank Big Data infrastructure (called “datapool”). The meaningful information was extracted and flattened in a single table. This task is particularly challenging because it is needed to identify events and instances from

the log file corresponding to the money transfer operations carried out by an employee from a bank centre and to connect those related to the same bank transfer. The heterogeneous nature of the log files, as saved in the CaixaBank datapool, makes this task even more difficult. There is a total of 969,351,155 events in the log data just for April 2019. These events are heterogeneous in nature and arise from mixing of disparate operations associated with services provided by the different types of bank offices. After a laborious table flattening and composition process, a table of 32 fields was obtained and then tokenised.

3.3 Enhanced control of customers to online banking

In this use case, we focused on analysing the mobile-to-mobile bank transfers executed via online banking (web and application). It focuses on the assessment that the controls applied to user authentication are adequately implemented (e.g., Strong Customer Authentication -SCA- by means of second-factor authentication) according to PSD2 regulation and depending on the context of the bank transfer. With that aim, we wanted to cluster a dataset collected from mobile-to-mobile transfers. Most of the information of this dataset does not need encryption because only a few fields were sensitive. The main objectives of the use case are to identify useful patterns of mobile-to-mobile bank transfers and enhance current cybersecurity mechanisms by identifying if there is a set of transactions in which the level of security in the authentication process should be increased.

4 I-BiDaaS solutions for the defined use cases

4.1. Analysis of relationships through IP addresses

4.1.1 Architecture

The architecture uses a traditional component-based architecture where the components communicate via a message queue (Universal Messaging component). This approach is important for a scalable and flexible hardware resource organisation. The architecture includes a batch and a stream processing subcases, complemented with the Universal Messaging component (mentioned earlier) for easing communication between components, an Orchestration layer that coordinates their interaction, and a Visualisation layer providing an extensible visualisation framework that helps in the data inspection and user interaction with the system. The Universal Messaging component uses a message queue system that allows easy, robust and concurrent communication between components. The Orchestration layer uses Docker for managing other components, which are all running individually as Docker containers. Figure 1 depicts the components of the architecture, as well as their interactions.

The batch processing subcase starts with the creation of a file of realistic synthetic data in SQLite format (TDF component⁵), which is then imported in a Cassandra Database which is specifically used for its distributed properties, and

⁵ TDF: <https://www.ibm.com/il-en/marketplace/infosphere-optim-test-data-fabrication>.

COMPSs [7] with Hecuba⁶ are used to run the analysis. In the streaming subcase, transactions of users are created and published via the Message Queuing Telemetry Transport (MQTT) protocol (Universal Messaging), and later an APAMA⁷ GPU-enabled data processing application loads the data analysis created in the batch subcase, and compares any data coming from the stream to it, generating a new message if there is a match. 3D data analysis through visualisation is also available via the Qbeast tool [8].

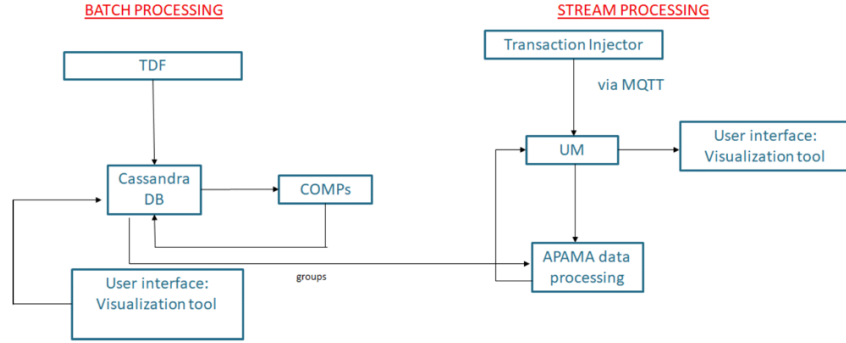


Figure 1: Updated data flow for the components of the Architecture

The rationale for using realistic synthetic data (TDF component in Figure 1) is that technology development and testing processes can be simplified and accelerated, before on in parallel with carrying out the processes of making real data available (e.g., a tedious data tokenisation process). The incorporation of realistic synthetic data is done with care and is subject to data quality assessment (see Section 4.1.2.) An operational-ready solution then replaces the realistic synthetic data with the corresponding real, tokenised data, as described in subsequent sections.

4.1.2 Data generation

In this first use case, we tried to evaluate the usage of fabricated data, which was created using TDF according to a set of rules defined by CaixaBank. The rules were refined several times in order to create realistic data for all different fields considering the format of the real data. It is difficult to distinguish a data sample from a field in the synthetic dataset and a sample from the same field in the real dataset. Some properties were difficult to model as constraint rules, e.g. the concrete time connectivity patterns that the real data follows, and thus they were not included in the specification of the synthetic dataset. Constraints for parameters which were not critical for the relationship analysis that was performed in the use case were sometimes relaxed as long as they allowed the synthetic dataset to remain valid for assessing that there exists the same percentage of relationships as in the real dataset.

⁶ Hecuba: <https://github.com/bsc-dd/hecuba>.

⁷ APAMA: https://www.softwareag.com/corporate/products/apama_webmethods/analytics/overview/default.asp.

4.1.3 Data analytics

The goal of the case is to find relations between people, given a set of connections to IP addresses, maximising the detection of close relations between users. This application has been implemented using COMPSs programming model and Hecuba as the data interface with the Cassandra database.

We have defined several parallel tasks, not only to exploit parallelism but also to benefit from the automatic detection of the dependencies from COMPSs. Using Cassandra to store the data allows us to delegate on the database the management of the global view of the data. This approach frees programmers from implementing an explicit synchronisation between those parallel tasks that modify the data structure. This way, removing the synchronisation points, we are able to maximise the parallelism degree of the application and thus the utilisation of the hardware resources. Notice that the interface for inserting data in Cassandra is asynchronous with the execution of the application, overlapping this way the data storing with the computation.

The approach to solve this implementation has been to define a clustering-based analysis of CaixaBank's IP address connections using a synthetic dataset. The purpose of the analysis is to provide additional modelling possibilities to this CaixaBank's use case. The obtained results should be understood relative to the fact that the data set utilised is synthetic, even though the initial feedback from CaixaBank about the usefulness of the developed process is positive, and the approach is promising. The data set contains 72.810 instances, with each instance containing the following attributes:

- *User ID* – representing a unique identification number for each user;
- *IP address* – representing the IP address of the connection of the user;
- *Date* – representing the date and the time of the connection made by the user;
- *Operation* – representing the code of the business operation made by the user;
- *Status* – representing the code of the status of the operation made by the user.

Initially, the dataset is transformed as follows: each user represents a sample, while each IP address represents a feature. In such a data matrix, the value in position (i, j) represents the number of times user i connected via IP address j . Such a dataset turns out to be extremely sparse. In order to tackle this problem and retain only meaningful data, the next pre-processing step is to drop all the IP addresses that were used by only one user (intuitively, such IP addresses represent home network, etc. and thus cannot be used to infer relationships between users). After dropping all such IP addresses, 1,075 distinct IP addresses remain from the initial 22,992 contained in the original dataset. Subsequently, we filter out the users which are not connected to any of the remaining IP addresses.

To infer relationships between users, we applied clustering algorithms. In particular, we used K-means [12] and DBSCAN [13], which are both available in the dislib library [10]. Additionally, we used the t-distributed Stochastic Neighbour

Embedding (t-SNE) method [11] to visualise the reduced dataset in 2D. The visualisation is presented in Figure 2.

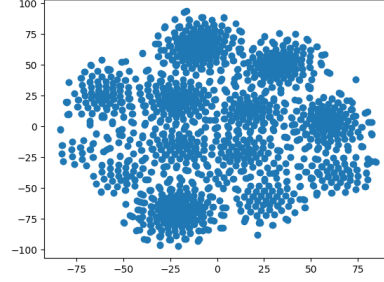


Figure 2. t-SNE 2D visualisation

Both K-means and DBSCAN offer some interesting hyperparameters. In particular, K-means allowed us the flexibility of setting the desired number of clusters. On the other hand, DBSCAN decides on the number of clusters internally while providing us with the parameters that represent the minimum number of samples in a neighbourhood for a point to be considered a core point, and the maximum distance between two samples for them to be considered as in the same neighbourhood. These parameters are to be set by an end-user based on experimentation and domain knowledge and are tuneable through the I-BiDaaS user interface.

Moreover, the evaluation of this use case was especially focused on analysing the validation of fabricated data for identifying patterns and number of connections. Therefore, a more advanced analysis with K-means and DBSCAN was done using both, the synthetic dataset and a tokenised version of a real dataset. The data tokenisation process included the encryption of all the fields of the dataset. The analysis performed over this dataset allowed the inference of conclusions and relationships in the real non-encrypted data.

4.1.4 Visualisations

The visualisation of the use case includes several graphic types. First, a graph shows the distribution of relationships detected based on their IP addresses (Figure 3a).



Figure 3 a) User groups per IP relationships



Figure 3 b) Real-Time relationship detection

Using these relationships, visualisation of real-time bank transfers in the form of a continuous stream of sender-receiver records is used to emulate real-time detection of possibly fraudulent transactions (Figure 3b). The visualisation utilises the previously detected relationships to display a graph of connected users so as to aid operators in determining possible relationships between users and decide whether further actions should be taken.

4.1.5 Results

Results obtained from both real tokenised data and the synthetic data using those algorithms showed that the majority of the clusters found were 2-point clusters, indicating a good similarity for this use case.

Additional evaluation process was performed to determine a specific utility score, i.e., the similarity of results of analyses from the synthetic data and the original data. The propensity mean-squared-error (pMSE) was used as a general measure of data utility to the specific case of synthetic data. As specific utility measures, we used various types of data analyses, confidence intervals overlap and standardised difference in summary statistics, which were combined with the general utility results.

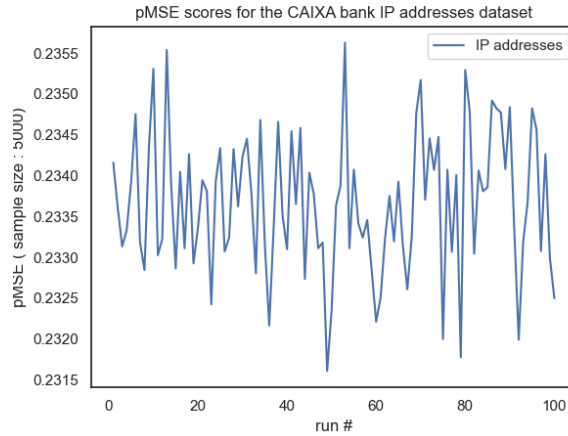


Figure 4. Results for 100 random sampling taken from the real and the synthetic data (5K datapoints each) and the pMSE calculated using a logistic model

Randomly sampling 5000 datapoints from the real and synthetic datasets, and using logistic regression to provide the probability for the label classification, we were able to show that the measured mean pMSE score for the ‘Analysis of relationships through IP addresses’ dataset is 0.234 with a standard deviation of 0.0008.

Those quantitative results showed that the fabricated data is objectively realistic to be used for testing the use case. However, the rule-generation process that involves the data fabrication through TDF can be complex and long in other cases in which the knowledge of the data is not complete or the extraction of rules through statistical analysis is not clear.

4.2 Advanced Analysis of bank transfer payment in financial terminal

4.2.1 Architecture

To run this use case, its architecture (i.e. the specific components of the I-BiDaaS general architecture) is the same as the one described in Section 4.1.1, focused on the batch processing part. Therefore, what essentially changes are the algorithms used for processing the data (i.e. the bank transfers conducted by employees on their financial terminals). These algorithms will be described in the next subsections.

4.2.2 Data analytics

This CaixaBank use case is focused on advanced analysis of bank transfers executed by employees on financial terminals to detect possible fraud, or any other potential anomalies that differ from the standard working procedure. The used dataset is composed of different attributes which record the different steps that the employee performs and other important data such as the account, client, or amount of money transferred. All the data is encrypted using the Dice Coefficient [14], which codifies the data without losing important information.

All data processing techniques, like the K-means, PCA (Principal Component Analysis) [15] and DBSCAN have been performed using the dislib library. Also, the data structure used by dislib has been modified to be stored on a Cassandra Database using the Hecuba library.

The received dataset must be pre-processed before using the data transformation techniques from dislib. First, the attributes which only contained the same value for all the registers have been deleted, as they do not give any relevant information. Also, all nulls and blank registers have been transformed into 0 values. Finally, for those categorical attributes, we transform the variable categories into columns (1, 0), a transformation known as one-hot encoding [16].

Due to the encoding transformations, the number of attributes has increased considerably from 89 to 501. This large amount of attributes made difficult to perform K-means, and for this reason, it was decided to apply a PCA transformation to reduce the number of dimensions to 3, to also be able to represent it graphically. Before applying the PCA transformation, and due to the differences in the magnitude of the attributes, we have standardised the data, using the *scikit-learn* method *StandardScaler*.

Finally, we have executed two different clustering algorithms: DBSCAN and K-means. As K-means requires the desired number of clusters as an input parameter, we have executed first DBSCAN and we have used the obtained number of clusters as the input parameter of K-means.

4.2.3 Visualisations

For this use case, a 3D graph of the data and detected anomalies has been developed. Users can select parts of the graph to focus on and can also extract the specific data samples that are included in the selection.



Figure 5 Visualisation of detected anomalies in 3-D graph

4.2.4 Results

Figure 6a shows the graphical representation of the clusters generated by DBSCAN in a three-dimensional space, where the third dimension of the PCA is shown as the Z-axis. The result of K-means can be examined in Figure 6b, we can observe that some values in the Z-axis are far away from the main cluster and thus, are potential anomalies in the data.

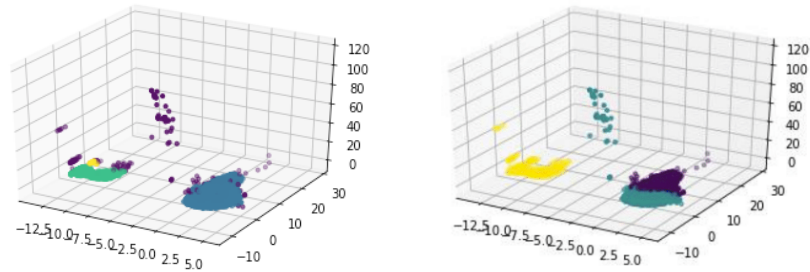


Figure 6: a) DBSCAN representation.

b) K-means representation.

The PCA reduced the attributes from 501 to 3, thus, it is difficult to understand which is the correlation between the resultant three dimensions and the 501 original attributes. In Figure 7, we have printed the mentioned correlation. We only show the first 84 because they are the most interesting with respect to the 3rd dimension of Z-axis. We can appreciate that this third dimension is heavily influenced by attributes from 64 to 82.

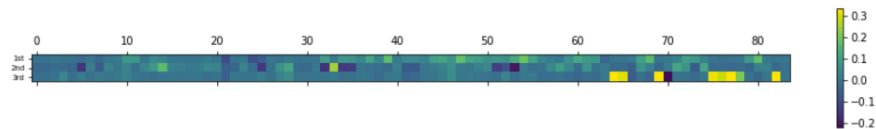


Figure 7: Heat-plot of the 84s most relevant attributes from the 501 original attributes.

4.3 Enhanced control of customers to online banking

4.3.1 Architecture

As in Section 4.2.1, the set of components used to analyse the bank transfers which were executed using online banking options (both web and mobile app) are the same as the ones described in Section 4.1.1, also with a main focus on the batch processing part, and selecting a different set of algorithms to analyse the data, as will be described in the following subsections.

4.3.2 Data analytics

Following the objective of the use case abovementioned in section 3.3, this use case tried to identify useful patterns of mobile-to-mobile bank transfers and enhance current cybersecurity mechanisms by identifying if there is a set of transactions in which the level of security in the authentication process should be increased. For that reason, we decided to analyse a dataset collecting the information of all the mobile-to-mobile bank transfers from clients within a month and work on non-supervised methods such as clustering. We needed to cluster on a categorical database so that most known algorithms lost efficacy. The first attempt was to apply a K-Means. However, since the vast majority of available variables were not numerical, calculating the distances for grouping in K-Means algorithm was no longer so simple (for example, if there are three types of enhanced authentication, the distance between them must be the same? Should it be greater since some of them are more restrictive than the others?) This type of question affects the result of the model and therefore, a transformation was made to the data. Therefore, we applied one-hot encoding [16]. This transformation allowed to eliminate the problems of calculating the distance between categories. Even so, the results were not satisfactory. Given the situation, a search/investigation process was carried out for an appropriate model for this case series. We find the *k*-modes library that includes algorithms to apply clustering on categorical data.

The *k*-modes algorithm [17] is basically the already known K-Means, but with some modification that allows us to work with categorical variables. The *k*-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimise the clustering cost function.

Once the algorithm has been decided, we must calculate the optimal number of clusters for our use case. For this, the method known as the *elbow* method is applied, which allows us to locate the optimal cluster as follows. We first define:

- *Distortion*: It is calculated as the average of the squared distances from the cluster centres of the respective clusters.
- *Inertia*: It is the sum of squared distances of samples to their closest cluster centre.

Then we iterated the values of *k* from 1 to 10 and calculate the values of distortion for each value of *k* and calculate the distortion and inertia for each value of *k* in the given range. The idea is to select the number of clusters that minimise inertia (separation between the components of the same cluster).

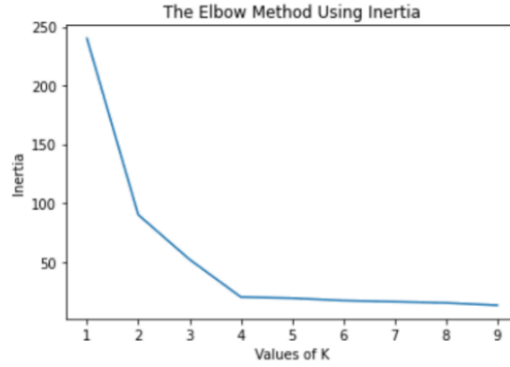


Figure 8: Number of clusters selection for ‘Enhanced Control of customers to Online Banking’

To determine the optimal number of clusters, we had to select the value of k at the ‘elbow’ in the point after which the distortion/inertia start decreasing in a linear fashion. Thus, for the given data, we conclude that the optimal number of clusters for the data is 4. Once we know the optimal number of clusters, we apply k -modes with $k = 4$ and analyse the results obtained.

4.3.3 Visualisations

A dynamically updated chart depicting the clusters in which the monitored transactions fall into was used for this use case. The number of clusters is automatically updated to reflect new ones being detected by the processing pipeline.



Figure 9. Sample of the I-BiDaaS graphical interface showing the identified clusters of incoming mobile-to-mobile bank transfers

4.3.4 Results

With this use case, I-BiDaaS allowed CaixaBank’s ‘Intermediate users’ and ‘Non-IT users’ to modify the number of clusters and run the algorithm over a selected dataset of transactions in a very fast and easy way. It was used for exploring clients’ mobile-to-mobile transaction patterns, identifying anomalies in the authentication methods and potential, allowing fast and visual analysis of the results in the platform (Figure 10).

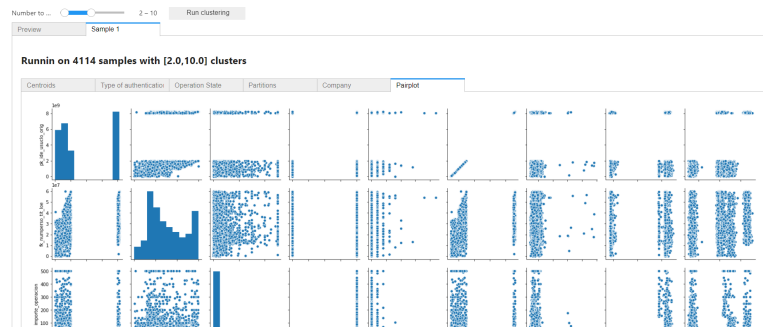


Figure 10: Sample of the ‘Enhanced Control of customers to Online Banking’ use case clustering results in the I-BiDaaS platform

Those results were checked with the Digital Security and Security Operation Centre (SOC) employees from CAIXA in order to correctly understand if the clustering algorithm applied allowed to identify potential errors in our automated authentication mechanisms in mobile-to-mobile bank transfers. The obtained clusters of entries were useful to identify the different mobile-to-mobile bank transfers patterns and reconsider the way we are selecting the authentication method to proceed with the transfer.

4.4 Relation with the BDVA reference model and the BDVA strategic research and innovation agenda

The described solution can be contextualised within the BDV reference model 0 defined in the BDV Strategic Research and Innovation Agenda (BDV SRIA) and contributes to the model in the following ways. Specifically, the work is relevant to the following BDV reference model horizontal concerns:

- **Data visualisation and user interaction:** We develop several advanced and interactive visualisation solutions applicable in the banking sector, as illustrated in Figures 3,5,9.
- **Data analytics:** We develop data analytics solutions for the three industrial use cases in the banking sector, as described in Section 4.1-4.3. While the solutions may not correspond to state-of-the art advances in algorithm development, they clearly contribute to revealing novel insights into how Big Data analytics can improve banking operations.
- **Data processing architectures:** We develop an architecture as shown in Figure 1 that is well-suited for banking applications where both batch analytics (e.g., analysing historical data) and streaming analytics (e.g., online processing of new transactions) are required. A novelty of the architecture is the incorporation of realistic synthetic data fabrication and the definition of scenarios of usefulness and quality assurance of the corresponding synthetic data.
- **Data protection:** We describe in Section 5 how data tokenisation and realistic synthetic data fabrication can be used in baking applications to allow for more agile development of Big Data analytics solutions.
- **Data management:** We present innovative ways for data management utilising efficient multidimensional indexing, as described in Section 4.3.

Regarding the BDV reference model vertical concerns, the work is relevant to the following:

- **Big Data Types and Semantics:** The work is mostly concerned with structured data, meta-data, and graph data. The work contributes to generation of realistic synthetic data from the corresponding domain-defined meta-data.
- **Cybersecurity:** The presented solutions that include data tokenisation correspond to novel best practice examples for securely sharing sensitive banking data outside bank premises.

Therefore, in relation with BDV SRIA, we contribute to the following technical priorities: Data protection; Data Processing Architectures; Data Analytics; and Data Visualisation and User Interaction.

5 Lessons learned, guidelines and recommendations

CaixaBank, as many entities in critical sectors, was initially very reluctant to use any Big Data storage or tool outside its premises. To overcome that barrier, the main goal of CaixaBank when enrolling in the I-BiDaaS project was to find an efficient way to perform Big Data analytics outside its premises. That would speed up the process of granting new external providers to access CaixaBank data (which usually encompasses a bureaucratic process that takes weeks or even a month). Additionally, CaixaBank wanted to be much more flexible in the generation of proof-of-concept (PoC) developments (i.e., to test the performance of new data analytics technologies to be integrated into its infrastructure). Usually, for any new technology testing, even a small test, if any hardware is needed to be arranged, it should be done through the infrastructure management subsidiary who will finally deploy it. Due to the size and level of complexity of the whole CaixaBank infrastructure and security assessment rigid processes, the deployment of it can take months.

For those reasons, CaixaBank wanted to find ways to bypass these processes without compromising the security of the entity and the privacy of its clients. General Data Protection Regulation (GDPR)⁸ really limits the usage of the bank customers' data, even if it is used for potential fraud detection and prevention and for enhancing the security of its customers' accounts. It can be used internally to apply certain security policies, but how to share this data with other stakeholders is still an issue. Furthermore, bank sector is strictly regulated, and National and European regulators are supervising all the security measures taken by the bank in order to provide a good level of security for the entity and, at the same time, maintain at any time the privacy of the customers. The current trend of externalising many services to the cloud also implies to establish a strict control of the location of the data and who has access to it for each migrated service.

The I-BiDaaS CaixaBank roadmap (Figure 11) had a turning-point, in which the entity 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

⁸ <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

and processes and evaluate tokenisation processes as an enterprise standard to extract data outside their premises, breaking both internal and external data silos.

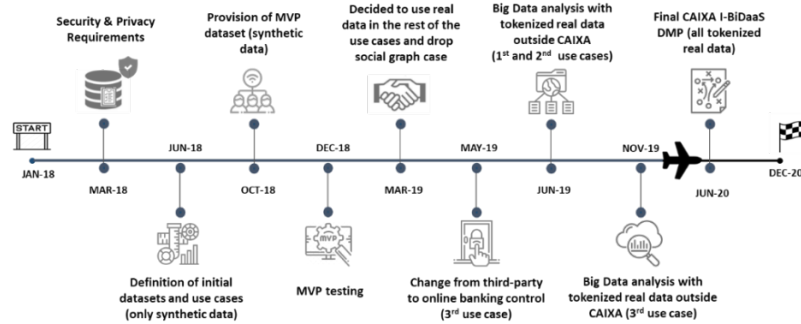


Figure 11. CaixaBank roadmap in I-BiDaaS project

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.

The other two use cases focused on how extremely sensitive data can be tokenised to extract real data for its usage outside CaixaBank premises. By tokenising, we mean encrypting the data and keeping the encryption keys in a secure data store that will always reside in CaixaBank facilities. This approach implied 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 tokenised 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), helping to 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.

The main benefits obtained by CaixaBank due its participation in I-BiDaaS (highlighted in Table 2) directly relate to the evaluation of the different requirements presented in Section 2 of this chapter (Table 1).

Table 2. Summary of the impact of the CaixaBank use cases studied 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.

We were able to speed up the implementation of Big Data analytics applications (R1), test algorithms outside CaixaBank premises (R2) and test new tools and algorithms without data privacy concerns by exploring and validating the usage of synthetic data and tokenised data (R3) in three different use cases, improving the efficiency in time and cost (R5, R6, R7) by means of skipping some data access procedures and being able to use new tools and algorithms in a much more agile way. User requirements regarding the availability of ‘Intermediate and Non-IT users’ to analyse and process the data of the use cases were also validated through several internal and external workshops⁹ in which the attendants from several departments of CaixaBank and other external entities (data scientists, business consultants, IT and Big Data managers) provided very positive feedback about the platform usability. Moreover, use cases 2 and 3, as mentioned previously, were also validated by the corresponding business processes employees, being able to extract the results by themselves.

Last but not least, it is important to highlight that those results should be applicable to any other financial entity that faces the same challenges and tries to overcome the limitations of data privacy regulation, the common lack of agility of large-scale on-premise Big Data infrastructures and very rigid but necessary security assessment procedures.

6 Conclusion

The digitalisation of the financial sector and the exploitation of the incredible amount of sensitive data collected and generated by the financial entities day-by-day makes their Big Data infrastructure very difficult to manage and to be agile integrating innovative solutions. I-BiDaaS integrated platform provided a solution to manage it in a much more accessible and friendly manner that approaches the Big Data analytics to the bank employees with less technical and data science knowledge. It also explored ways to reduce the friction between data privacy regulation and the exploitation of the sensitive collected data for several purposes, showcasing it for several use cases on enhancing the entity cybersecurity and preventing fraud to their clients.

⁹ <https://www.ibidaas.eu/blog/%e2%80%9cI-BiDaaS-Application-to-the-Financial-Sector%e2%80%9d-Workshop/>

Acknowledgements

The research presented in this book chapter was undertaken in the framework of the I-BiDaaS project (“Industrial-Driven Big Data as a Self-Service Solution”) funded by the Horizon 2020 Programme under Grant Agreement 780787.

References

- [1] Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). “Euro-pean Big Data Value Strategic Research & Innovation Agenda”.
- [2] Passlick J., Lebek, B. and Breitner, M.H. “A Self-Service Supporting Business Intelligence and Big Data Analytics Architecture”, 13th International Conference on Wirtschaftsinformatik, St. Gallen, Switzerland. February 12-15, 2017.
- [3] B Arruda, D.: Requirements engineering in the context of big data applications. SIG-SOFT Softw. Eng. Notes 43(1), 1-6. Mar 2018.
- [4] Horkoff, J., A.F.C.E.: Goal-oriented requirements engineering: an extended systematic mapping study. Requirements Eng 24(2019), 133-160. 2019.
- [5] NIST Big Data Public Working Group: Use Cases Requirements Subgroup: National Institute of Standards and Technology (NIST) “Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements”. Tech. rep., National Institute of Standards and Technology, Special Publication 1500-3. 2015.
- [6] I-BiDaaS Consortium. “D1.3: Positioning of I-BiDaaS”, http://www.ibidaas.eu/sites/default/_les/docs/Ibidaas-d1.3.pdf. 2018.
- [7] Badia, R. M., J. Conejero, C. Diaz, J. Ejarque, D. Lezzi, F. Lordan, C. Ramon-Cortes, and R. Sirvent, “COMP Superscalar, an interoperable programming framework”, SoftwareX, Volumes 3–4, December 2015, Pages 32–36, DOI: 10.1016/j.softx.2015.10.004.
- [8] Cugnasco, C., Calmet, H., Santamaria, P., Sirvent, R., Eguzkitza, A. B., Houzeaux, G., Becerra, Y., Torres, J. & Labarta, J.. “The OTree: Multidimensional Indexing with efficient data Sampling for HPC”. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 433-440). IEEE. December 2019.
- [9] Arapakis, I. et al. "Towards specification of a software architecture for cross-sectoral big data applications." 2019 IEEE World Congress on Services (SERVICES). Vol. 2642. IEEE, 2019.
- [10] Álvarez Cid-Fuentes, J., Solà, S., Álvarez, P., Castro-Ginard, A. and Badia, R. M. “dislib: Large Scale High Performance Machine Learning in Python,” in Proceedings of the 15th International Conference on eScience, 2019, pp. 96-105. 2019.
- [11] Hinton, Geoffrey E., and Sam Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems 15: 857-864. 2002
- [12] Bock, H-H. "Clustering methods: a history of k-means algorithms." Selected contributions in data analysis and classification. Springer, Berlin, Heidelberg. 161-172. 2017.
- [13] Martin E., et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.
- [14] Jimenez, S., Becerra, C. and Gelbukh, A. “Soft cardinality: A parameterized similarity function for text comparison”. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 449-453). 2012.
- [15] Pearson, K.. "LIH. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2.11 (1901): 559-572.
- [16] Okada, S., Ohzeki, M. and Taguchi, S. “Efficient partition of integer optimization problems with one-hot encoding”. Scientific reports, 9(1), pp.1-12. 2019.
- [17] Huang, Z. “Extensions to the k-modes algorithm for clustering large data sets with categorical values”, Data Mining and Knowledge Discovery 2(3), pp. 283-304, 1998.