



Horizon 2020 Program (2014-2020)

Big data PPP

Research addressing main technology challenges of the data economy



Industrial-Driven Big Data as a Self-Service Solution

D6.4: Experiments implementation[†]

Abstract: This deliverable provides a detailed description of the experiments' implementation following the line of D6.2 [1] and taking into consideration the progress of the work done from M19 to M32 concerning the experimental protocol alignment of task 6.1, the implementation of the real-life industrial experiments of task 6.2 and the evaluation and impact analysis of task 6.3. Each experiment is described in terms of generated datasets, workflow and goals, by reporting the experimental indicators and associated metrics for all experiments. Furthermore, the deliverable reports on the impact analysis and external stakeholders' involvement during the progress of the project from M23 to M32.

Contractual Date of Delivery	31/08/2020
Actual Date of Delivery	31/08/2020
Deliverable Security Class	Public
Editor	<i>Giuseppe Danilo Spennacchio (CRF)</i>
Contributors	All I-BiDaaS partners
Quality Assurance	<i>Leonidas Kallipolitis (AEGIS)</i> <i>Dusan Jackovetic (UNSPMF)</i> <i>Kostas Lampropoulos (FORTH)</i>

[†] The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780787.

The I-BiDaas Consortium

Foundation for Research and Technology – Hellas (FORTH)	Coordinator	Greece
Barcelona Supercomputing Center (BSC)	Principal Contractor	Spain
IBM Israel – Science and Technology LTD (IBM)	Principal Contractor	Israel
Centro Ricerche FIAT (FCA/CRF)	Principal Contractor	Italy
Software AG (SAG)	Principal Contractor	Germany
Caixabank S.A. (CAIXA)	Principal Contractor	Spain
University of Manchester (UNIMAN)	Principal Contractor	United Kingdom
Ecole Nationale des Ponts et Chaussees (ENPC)	Principal Contractor	France
ATOS Spain S.A. (ATOS)	Principal Contractor	Spain
Aegis IT Research LTD (AEGIS)	Principal Contractor	United Kingdom
Information Technology for Market Leadership (ITML)	Principal Contractor	Greece
University of Novi Sad Faculty of Sciences (UNSPMF)	Principal Contractor	Serbia
Telefonica Investigation y Desarrollo S.A. (TID)	Principal Contractor	Spain

Document Revisions & Quality Assurance

Internal Reviewers

1. *Leonidas Kallipolitis (AEGIS)*
2. *Dusan Jakovetic (UNSPMF)*
3. *Kostas Lampropoulos (FORTH)*

Revisions

Version	Date	By	Overview
0.09	31/08/2020	G.D. Spennacchio	Final version
0.08	28/08/2020	Quality assurance, Internal reviewers	Comments by AEGIS and UNSPMF on final version
0.07	26/08/2020	G.D. Spennacchio	Final draft
0.06	24/08/2020	Quality assurance, Internal reviewers	Comments by AEGIS on first draft
0.05	12/08/2020	Quality assurance, Internal reviewers	Comments by UNSPMF on first draft
0.04	10/08/2020	Contributors	First draft
0.03	27/07/2020	G.D. Spennacchio	First draft
0.02	05/05/2020	Quality assurance, Internal reviewers	Comments on the ToC
0.01	21/04/2020	G.D. Spennacchio	ToC

Table of Contents

LIST OF TABLES.....	6
LIST OF FIGURES.....	7
LIST OF ABBREVIATIONS.....	9
EXECUTIVE SUMMARY.....	10
1 INTRODUCTION.....	11
2 EXPERIMENTAL PROTOCOL ALIGNMENT	12
3 IMPLEMENTATION AND OPERATION OF REAL-LIFE INDUSTRIAL EXPERIMENTS.....	13
3.1 OVERVIEW	13
3.2 GENERATED DATASETS.....	14
3.2.1 <i>Fabricated datasets</i>	14
3.2.2 <i>TID datasets</i>	15
3.2.3 <i>CAIXA datasets</i>	17
3.2.4 <i>CRF datasets</i>	19
3.2.5 <i>Generic use case datasets</i>	22
3.3 INDUSTRIAL EXPERIMENTS IMPLEMENTATION.....	23
3.3.1 <i>Experimental workflow</i>	23
3.3.2 <i>Telecommunication experiments</i>	24
3.3.2.1 <i>Accurate location prediction with high traffic and visibility</i>	24
3.3.2.2 <i>Optimization of Placement of Telecommunication Equipment</i>	26
3.3.2.3 <i>QoS in Call Centres</i>	28
3.3.3 <i>Banking experiments</i>	30
3.3.3.1 <i>Analysis of relationships through IP addresses use case</i>	31
3.3.3.2 <i>Advanced Analysis of bank transfer payment in financial terminal</i>	39
3.3.3.3 <i>Enhanced control of customers to online banking</i>	41
3.3.3.4 <i>Data Encryption</i>	44
3.3.4 <i>Manufacturing experiments</i>	48
3.3.4.1 <i>Maintenance and Monitoring of production assets</i>	48
3.3.4.2 <i>Production Process of Aluminium die-casting</i>	52
3.3.5 <i>Generic experiments</i>	55
3.3.6 <i>Cross-sectorial experiments</i>	57
4 EXPERIMENTAL EVALUATION.....	58
4.1 OVERVIEW	58
4.2 DATA QUALITY EVALUATION	58
4.2.1 <i>Data quality from the perspective of assessing algorithm scalability</i>	58
4.2.2 <i>Specific and general utility</i>	59
4.3 THE I-BiDaaS INTEGRATED SOLUTION AND ARCHITECTURE IMPLEMENTATION	62
4.4 EXPERIMENTS VERIFICATION AND VALIDATION	65
4.5 QUANTITATIVE EVALUATION	66
4.5.1 <i>Individual parts evaluation</i>	66
4.5.2 <i>Overall I-BiDaaS solution evaluation</i>	74
4.5.3 <i>Tests in relation to I-BiDaaS industry validated benchmarks</i>	75
4.6 QUALITATIVE EVALUATION	75
4.6.1 <i>Telecommunication experiments</i>	75
4.6.2 <i>Banking experiments</i>	77
4.6.3 <i>Manufacturing experiments</i>	79
4.6.4 <i>Generic use case experiments</i>	80
4.6.5 <i>High-level non-functional requirements evaluation</i>	82
5 IMPACT ANALYSIS.....	83
5.1 PROGRESS REPORT	83
5.2 FEEDBACK FROM EXTERNAL STAKEHOLDERS.....	90
5.3 EXPLOITATION AND POTENTIAL COMMERCIALIZATION	93
6 CONCLUSION.....	94

7	REFERENCES.....	95
8	APPENDIX.....	97

List of Tables

Table 1: Datasets Description.....	13
Table 2: Structure of the dataset for the SCADA data.....	20
Table 3: Structure of the dataset for MES data.....	20
Table 4: Structure of the synthetic dataset for the Production process of Aluminium die-casting.....	21
Table 5: Structure of the real anonymised dataset for the Production process of Aluminium die-casting	21
Table 6: Coordinates dataset, used for K-Means clustering algorithm	22
Table 7: Coordinates dataset with labels, used for K-Means evaluation algorithm.....	22
Table 8: Accurate Location Prediction with High Traffic and Visibility.....	25
Table 9: Optimization of Placement of Telecommunication Equipment.....	27
Table 10: QoS in Call Centres	30
Table 11: Silhouette score for K-Means clustering.....	35
Table 12: Silhouette score and number of clusters found for DBSCAN clustering.....	35
Table 13: Sizes of communities and their frequencies.....	37
Table 14: Analysis of relationships through IP address	38
Table 15: Advanced Analysis of bank transfer payment in financial terminal.....	40
Table 16: Enhance control of customers to online banking	43
Table 17: Maintenance and monitoring of production assets	51
Table 18: Production process of aluminium die-casting	55
Table 19: Experiment definition for end-to-end I-BiDaaS platform in self-service mode.....	56
Table 20: Experiment definition for end-to-end I-BiDaaS platform in expert mode.....	56
Table 21: Component evaluation	66
Table 22: Qbeast speedup after few ReadOptimizations.....	72
Table 23. Overall I-BiDaaS prototype evaluation results.....	74
Table 24: Experiment #1 - Accurate Location Prediction with High Traffic and Visibility	76
Table 25: Experiment #2 - Optimization of Placement of Telecommunication Equipment	76
Table 26: Experiment #3 - QoS in Call Centres	77
Table 27: Experiment #4 - Enhance control of customers to online banking.....	78
Table 28: Experiment #5 - Advanced Analysis of bank transfer payment in financial terminal	78
Table 29: Experiment #6 - Analysis of relationships through IP address	79
Table 30: Experiment #7 - Maintenance and monitoring of production assets	79
Table 31: Experiment #8 - Production process of aluminium casting	80
Table 32: Experiment #9 - Experiment for end-to-end I-BiDaaS platform in self-service mode.....	80
Table 33: Experiment #10 - Experiment for end-to-end I-BiDaaS platform in expert mode.....	81
Table 34: Progress with regards to I-BiDaaS KPIs.....	84
Table 35: Explanation of the status and the next steps of the I-BiDaaS KPIs.....	85
Table 36: I-BiDaaS Tools & Technologies	89
Table 37: List of I-BiDaaS Excellent Innovations as accepted from EU Innovation Radar	89

List of Figures

Figure 1: Thermal image in jpg format	22
Figure 2: Thermal data in csv format	22
Figure 3: Mobility Data Locations (normalised by the mean).....	25
Figure 4: Antenna KPIs - Hotspot Prediction (Live Streaming).....	27
Figure 5: Example execution in TID server of the sentiment analysis tool.....	28
Figure 6: Advanced Visualization Toolkit (AVT) supporting scalable data visualisation.....	29
Figure 7: Example of Live Stats & Daily AVG (Time Window: 1 hour)	29
Figure 8: Analysis of relationships through IP address use case visualisation in I-BiDaaS platform... <td>32</td>	32
Figure 9: Visualisation of the found relationships between bank transfers' senders and receivers.	32
Figure 10: Graph visualisation of the clients' relationships through IP address.....	33
Figure 11: t-SNE projection of the dataset in 2D.....	34
Figure 12: Sample of the 'Advanced analysis of bank transfer payment in financial terminal' use case in I-BiDaaS Expert Mode visualisation	39
Figure 13: 'Advanced analysis of bank transfer payment in financial terminal' use case analysed with DataRobot.....	40
Figure 14: Number of clusters selection for 'Enhanced Control of customers to Online Banking'	42
Figure 15: Sample of the 'Enhanced Control of customers to Online Banking' use case clustering results in the I-BiDaaS platform	43
Figure 16: Bloom-filtered arrays resulted from two similar texts.....	46
Figure 17: Result of street name similarity after being encrypted with Bloom-filters.....	47
Figure 18: Anomalous number per day for the selected sensor and month.....	49
Figure 19: Sensor's mean value and anomalous trend for the selected sensor and day	50
Figure 20: Sensor calendar	50
Figure 21: Categories of sensors.....	50
Figure 22: Sensors of acceleration	51
Figure 23: Dynamic diagram for the 'Production process of Aluminium die-casting' use case	53
Figure 24: Parameters trend.....	53
Figure 25: Classification level	54
Figure 26: Classification levels versus engine produced	54
Figure 27: Execution time versus number of cores for the real (tokenized) dataset (red line) and the TDF-generated synthetic dataset (green line) for the CAIXA's use case 'Analysis of relationships through IP addresses'	58
Figure 28: Data attributes analysed for single and multi-columns	60
Figure 29: An excerpt of a single column (PM2)	60
Figure 30: Results from 100 random sampling taken from the real and the synthetic data (5000 datapoints of each) with the pMSE calculated from a logistic model	61
Figure 31: Login to I-BiDaaS platform.....	62
Figure 32: Different modes supported by the I-BiDaaS platform.....	62
Figure 33: Expert mode	62
Figure 34: Self- service mode	63
Figure 35: Co-Develop mode	63
Figure 36: The I-BiDaaS architecture	64
Figure 37: System architecture: The orchestrator and the docker swarm	64
Figure 38: System architecture: The orchestrator and APAMA analytics	65
Figure 39: Thousands of IOPS of Cassandra with 2 replicas vs Qbeast	71
Figure 40: Net I/O time for 1000 steps with different backends	71
Figure 41: Response time: Qbeast vs. PostgreSQL	72
Figure 42: Response time K-Means.....	73
Figure 43: Response time PCA.....	73
Figure 44: Response time KNN.....	74
Figure 45: General questions webinar results: a. CaixaBank; b. Telefonica; c. CRF	91
Figure 46: Webinar results specific questionnaire: a. Telefonica; b. CRF	92

Figure 47: CaixaBank Workshop results..... 92

List of Abbreviations

ADMM	Alternating Directions Method of Multipliers
AMQP	Advanced Message Queueing Protocol
ASR	Automatic Speech Recognition
ATM	Automated Teller Machine
BDV	Big Data Value
CC	Call Centre
COMPSS	COMP Superscalar
CSI	Customer Satisfaction Index
CTM	Compressed Triangle Mesh
DL	Deep Learning
EBDVF	European Big Data Value Forum
FCR	First Call Resolution
FHE	Fully Homomorphic Encryption
FML	Federated Machine Learning
IOPs	Input/output operations per second
JMS	Java Messaging Service
JPH	Job per Hour
KPI	Key Performance Indicator
ML	Machine Learning
MNOs	Mobile Network Operators
MPI	Message Passing Interface
MQTT	Message Queueing Telemetry Transport
MVP	Minimum Viable Product
NLP	Natural Language Processing
NN	Neural Network
OEE	Overall Equipment Effectiveness
PPP	Public-Private Partnership
PSD2	Payment Services Directive 2
QoS	Quality of Service
RoI	Return on Investment
RTTM	Rich Transcription Time Marked
SCA	Strong Customer Authentication
SOC	Security Operation Centre
TDF	Test Data Fabrication, formerly known as DFP
TN/TP	True Negative / True Positive
t-SNE	t-Distributed Stochastic Neighbour Embedding
UM	Universal Messaging

Executive Summary

This deliverable describes the implementation and operation of real-life industrial experiments from the telecommunication, banking and manufacturing industries in order to demonstrate how the I-BiDaaS solution has been applied in real-world environments. The I-BiDaaS project, funded by the Horizon 2020 Programme under Grant Agreement 780787, 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. To this end, the project developed an integrated platform for processing and extracting actionable knowledge from big data that includes: 1) data ingestion from various data sources and its preparation; 2) fabrication of realistic synthetic data for experimentation and testing; 3) batch and streaming analytics; and 4) simple, intuitive, and effective visualization and interaction capabilities for the end-users.

All activities have been aligned with the I-BiDaaS experimental protocol in order to ensure a smooth and adequate running of the operational experiments in alignment with the business objectives identified by the industrial partners for all defined use cases. Further revisions were taken into account regarding the real-life industrial experiments during the implementation phase and the progress in the design of the I-BiDaaS platform and associated technologies.

In more detail, the deliverable reports on the description of the different types of data provided, the generated datasets, the experimental workflow and the quality data evaluation for all use cases exploited during the experimental process and integrated in the I-BiDaaS platform. The integration was performed by ensuring secure data management through the anonymization and encryption of the data. Several types of programming languages and advanced visualization tools have been used to develop a platform easy-to-use for all experiments. Furthermore, the different end-user categories for each sector were detailed to characterize the usage of the platform and two additional generic use cases for expert and non-expert users have been developed to offer a more generic solution beyond the I-BiDaaS use cases, taking into consideration that the solution depends not only on the type and amount of data, but also on the type of different potential end-users.

At last, the deliverable describes the progress about the impact analysis, by continuing the work started in D6.2 [1] and reported in D.6.3 [2] with respect to the expected project innovation and achievements and provides the external stakeholders' feedback collected during the progress of the project.

1 Introduction

This deliverable continues the work started in the D6.2 [1] and reports the detailed description of the results of '*WP6. Real – life industrial and operational experiments*'. All experiments were executed according to the experimental protocol alignment (Task 6.1), were implemented and operated through three different real scenarios belonging to the telecommunication, financial and manufacturing sectors (Task 6.2) and tests were defined to determine the efficiency, operability, usability, robustness, performance, privacy awareness and costs of the real experiments and the impact analysis (Task 6.3). Each experiment was defined within the project in terms of data gathering, datasets implementation, analysis, integration and explanation of experimental results.

Synthetic and real anonymised data have been provided, generated and processed. The methods, developed in WP2 '*Data curation, ingestion and pre-processing*', have been used to aggregate, pre-process, manage and synthesize different types of data in both batch and real time. Batch and stream processing, described in detail in deliverables submitted under WP3 '*Batch processing innovative technologies for rapidly increasing historical data*' and WP4 '*Distributed analytics over extremely large numbers of high volume streams*', have been performed in WP6 activities in order to take into account all aspects, which may occur in real-world environments, such as cases that require a deeper analysis of large amounts of data, collected over a period of time (batch) or those that require velocity and agility for the events that we need to monitor in real or near real-time (streaming). Operational experiments and trials have been carried out using the I-BiDaaS solution within an interactive process between data providers and I-BiDaaS analysts and technologists.

All KPIs have been updated, taking into consideration the progress made from M19 to M32 for each experiment. I-BiDaaS solution aims to demonstrate in a realistic, measurable, and replicable way the effects that Big Data have on different real scenarios. The project is designed to use and validate Big Data by increasing operational efficiency and fostering new business models. Within I-BiDaaS, the capability of Big Data innovations to develop more efficient solutions is demonstrated by executing, evaluating and validating real operational scenarios (Pilots) belonging to the Telecommunication, Financial and Manufacturing domains.

Finally, this deliverable reports on the progress of the impact analysis with respect to the expected project level innovation and achievements. Furthermore, it provides a description of the activities that involved external stakeholders, who have expertise, experience or interest in Big Data analytics, in the evaluation process.

The rest of the document is structured as follows. Section 2 reports on the experimental protocol alignment based on the incremental and iterative nature of the I-BiDaaS solution applied to real life experiments. Section 3 provides a detailed description of the implemented datasets and the experimental workflow of the industrial experiments' implementation for each use case. Section 4 describes the experimental evaluation in terms of data quality, I-BiDaaS solution, architecture implementation, experiments verification and validation. Section 5 discusses the impact analysis and provides an overview of external stakeholders' involvement activities. Section 6 concludes with a summary of the results of WP6 achieved from M19 to M32.

2 Experimental protocol alignment

The experimental protocol alignment process aims to refine and, if necessary, revise the outcomes of the initial experiment's definition phase according to the I-BiDaaS experimental protocol (see D1.3 [3]) and to fine-tune the details of the industrial experiments to assure that the designed experiments will validate both business and technical requirements.

A detailed description of the alignment process, together with an overview of the initial alignment of the I-BiDaaS experiments, has been reported in deliverable D6.3 [2]. Due to the incremental and iterative nature of the I-BiDaaS experimental protocol, further alignment was required in order to reflect (a) revisions of the industrial use cases definitions and (b) progress in the design of the I-BiDaaS platform and associated technology characteristics.

This resulted in the revision of the definition of the I-BiDaaS experiments in terms of:

- a) The experiment's goals, data sets, analytics type, workflow and participants.
- b) The experimental indicators and associated metrics to be measured during the experiment.

Overall, ten experiments have been defined (reported in section 3). Eight of them are real-life industrial experiments that address real problems in the telecommunication, financial and manufacturing sectors. These experiments reflect the specific requirements of the project industrial partners and correspond to the 'Co-Develop mode' of operation of the I-BiDaaS platform, whereby end-users receive support and guidance from I-BiDaaS members in order to customize the analytics pipeline and enhance the visualization of the experiment results.

End-users, in this case, may fall in the following general categories:

- Data Providers and/or Data Consumers: Business users who introduce new data or information feeds into the platform, and/or use the Big Data analytics services and results. These, depending on the experiment, include data analysts, quality assurance and control managers, financial administrators, infrastructure engineers.
- Other stakeholders (e.g. IT security personnel): these are not end users per-se, however, they are evaluators or administrators of the platform.

See Section 3.3 for a more fine-grained description of users per different sector.

In addition, two generic experiments were defined, aiming to evaluate how the I-BiDaaS solution can be applied generally, reflecting the requirements of two generic user categories:

- Non-expert users: correspond to business users (data analysts) of the platform in 'self-service' mode. Such users understand the basic concepts of data analytics, machine learning and statistics.
- Expert users: correspond to Big Data developers that use the platform in 'expert mode'. They are able to develop data analysis applications in COMPSs or at least 'pure' Python (expert user).

3 Implementation and operation of real-life industrial experiments

3.1 Overview

International organizations and different competitive scenarios have been selected for developing, implementing and evaluating real-world industrial experiments in the EU H2020 I-BiDaaS project. Three data providers, namely TID (Telefonica I+D), CAIXA (CaixaBank) and CRF (Centro Ricerche FIAT), belonging to the telecommunication, banking and manufacturing sectors respectively, have defined eight real-world, industry-lead experiments where I-BiDaaS solution is being tangibly validated. In addition to the real-life experiments, two generic use cases have also been defined, considering potential non-expert/expert end-users and inputs for cross-sectorial experiment have been provided.

Each experiment has been carried out by utilizing different datasets (synthetic/real data) or processing type (batch/stream). Table 1 provides an overview of the datasets generated from all three industrial partners and detailed in Section 3.2, for each one of them.

Table 1: Datasets Decscription

No.	Use Case	I-BiDaaS dataset	Dataset Preparation Status	Data Provider
1	Accurate location prediction with high traffic and visibility	Anonymized TID mobility data Synthetic TID mobility data	Generated – 100% Generated - 100%	TID
2	Optimization of placement of telecommunication equipment	Anonymized TID mobility data Synthetic TID mobility data	Generated - 100% Generated - 100%	TID
3	Quality of Service in Call Centres	Anonymized call centre data Synthetic call centre data	Generated - 100% Generated - 100%	TID
4	Enhance control of customers to online banking access	Online banking (mobile to mobile bank transfer) tokenized data	Generated - 100%	CAIXA
5	Advanced analysis of bank transfer payment in financial terminal	Bank transfer tokenized data	Generated - 100%	CAIXA
6	Analysis of relationships through IP addresses	Synthetic IP address data Tokenized IP address data	Generated - 100% Generated - 100%	CAIXA
-	Building of a social graph	Synthetic Social graph data	DROPPED	CAIXA
7	Maintenance and monitoring of production assets	Anonymized SCADA data Anonymized MES data	Generated - 100 % Generated - 100%	CRF
8	Production process of aluminium die-casting	Anonymized Aluminium die-casting data (process and thermal data) Synthetic Aluminium die-casting data	Generated - 100% Generated - 100%	CRF

9	Experiment for end-to-end I-BiDaaS platform in self-service mode	Synthetic datasets used for proof of concept	Generated – 100%	I-BiDaaS consortium	
10	Experiment for end-to-end I-BiDaaS platform in expert mode	Bank transfer tokenized data (same as no.5)	Generated – 100%	CAIXA	
11	Cross-sectorial experiments	This experimentation was thoroughly investigated. However, a concrete experiment was not pursued in detail due to inherent constraints. See subsection 3.3.6 for details.			

The telecommunications industry collects massive amounts of data that act as the catalyst for business improvement. TID tested three use cases in order to improve the customer experience by employing advanced Machine Learning techniques. Part of the effort of improving the customer experience is focused around the employment of voice activated bots that help the users accomplish tasks related to the network configuration and operation. For the Telecommunication use cases, anonymized/synthetic data are analysed to predict changes in the number of connected mobile phone users per sector and the Customer Satisfaction Index (CSI).

CAIXA, as a representative of the financial sector in the project, tested three use cases that revolve around the huge amount of data collected by the different sources (ATMs, online banking services, employees' workstations, external providers' activity, network devices, etc.). For financial use cases, data analysts used synthetic/tokenized data for developing algorithms and tool performance testing or proof-of-concepts' validation skipping the strict security and privacy internal validation procedures of CAIXA.

The manufacturing industry, represented in the I-BiDaaS project by CRF, generates a large amount of heterogeneous data from various devices, systems and applications that enable manufacturers to develop new methodologies for the Big Data era. CRF is testing two use cases in order to demonstrate the ability to exploit I-BiDaaS solution to take profit of the near real-time shop floor data and to apply sophisticated statistical assessments. For the manufacturing use cases, data analysts used real or anonymized data, retrieved from the production lines, used for continuous improvement of algorithms in order to avoid cost breakdown, micro or macro stoppages and decrease of quality level. Unnecessary actions, such as preventive or planned maintenance, retooling, refurbishing, or repair of products, will be drastically reduced.

Finally, experiments for end-to-end I-BiDaaS platform either in self-service mode or in expert mode have been defined to provide a service with functional solutions for non-expert/expert end-users who want to optimize the performance and efficiency of their businesses. Relevant dataset examples have also been provided for these generic experiments.

3.2 Generated datasets

In the following subsections, the synthetic, real encrypted and anonymized datasets are described.

3.2.1 Fabricated datasets

D2.1 [4] provides a detailed specification of the I-BiDaaS platform's heterogeneous datasets of the Telecommunication, Financial and Manufacturing targeted domains. More specifically, section 3 '**Definitions of the data assets**' of D2.1 [4] defines the processes and activities related to the organization and integration of the data collected for the various use cases and provides

a clear definition of the I-BiDaaS data assets, nature and formats. Each use case scenario contains an overview, a high-level description of the data (i.e. the origin of data, its content, type, format, storage, consumption, volume, velocity, variety, variability, volatility, consistency, sensitivity and privacy) and the concrete structure (e.g. schemas, tables, columns, attributes, columns semantics, columns statistical properties, constraints etc.).

TDF was used to generate datasets for the following uses cases:

- Analysis of relationships through IP addresses – see D2.1 [4] Section 3.6 ‘*Analysis of relationships through IP addresses*’.
- Production process of aluminium die-casting – see D2.1 [4] Section 3.9 ‘*Production process of aluminium casting*’.
- Accurate location prediction with high traffic and visibility and & Optimization of placement of telecommunication equipment – see D2.5 [5] Section 4.1 ‘*Data fabrication via simulation*’.

Synthetic data was fabricated on an I-BiDaaS dedicated Virtual Machine into PostgreSQL DB, SQLite and csv files. TDF projects were defined, and the data was successfully generated, however, changes in the fabrication approach were made (see section 4.1 ‘*Data fabrication via simulation*’ of D2.5 [5]) due to inaccurate results in the TID use case.

3.2.2 TID datasets

The datasets (synthetic or otherwise) provided by TID address three representative and relevant use cases:

- Accurate location prediction with high traffic and visibility
- Optimization of placement of telecommunication equipment
- Quality of service in Call Centres

In all three uses cases, anonymised Big Data has been made available to I-BiDaaS technologists and analysts in order to assess the quality of the call centre services, e.g. by providing transcripts of customer service phone calls, or aggregated and anonymised mobility and antenna logs, in order to perform predictions on user movements.

More specifically, for the ‘*Accurate location prediction with high traffic and visibility*’ use case, TID made available a dataset that consists of anonymous traces collected from a large European cellular network provider. Each trace is a time series of mobile events that contain the encrypted user identifier, a timestamp, and the location of the associated base station. The base stations have varied coverage (between ~100 m to tens of km) depending on deployment density and radio propagation characteristics like obstacles, hills, or mountains. The expected user displacement in urban areas is smaller than in rural areas and can reach as low as 70 m.

A mobile event is generated every time a mobile device:

- activates/deactivates in the network
- makes/receives a call
- sends/receives an SMS
- moves from one location area code to another
- changes from one technology to another
- requests access to data (2G/3G) or requests a high-speed data channel (4G)
- is actively pinged by the network if no other event is registered for 2 hours

More specifically, the dataset consists of approximately 120K traces X N_i events per user i , divided into four-hour periods (yielding 186 points in total). The first field is the hashed user identifier (UID); fields 2-5 are aggregated statistics related to: 1) distance traversed by the user, 2) time connected to cell sites, 3) from 6th field onwards, we have tuples of antenna ID and amount of time connected to that antenna. The length of each time series varies since, for example, if a user has moved a lot, the user would have connected to more number of antennas.

The dataset was provided in JSON format, where every antenna is a JSON object containing its time series for the period. The values in the time series represent the number of users using the antenna at some point in time, so the values are strictly positive. Every time series was split into training and validation sets; then the time series are fitted and tested with the respective models.

For the '*Optimization of placement of telecommunication equipment*' use case, TID aggregated 2G, 3G and 4G feeds. Mobile Network Operators (MNOs) continuously collect various Key Performance Indicators (KPIs), such as coverage monitoring, and voice/data service metrics, about each radio sector. Such antenna KPIs are one of the key information for MNOs to understand network performance, and are used as input for network management, planning, and optimization. The employed indicators correspond to 2G, 3G and 4G sectors and can be grouped into the following categories:

- coverage (e.g., radio interference, noise level, power characteristics)
- accessibility (e.g., success establishing a voice or data channel, paging success, allocation of high-speed data channels)
- retain ability (e.g., fraction of abnormally dropped channels)
- mobility (e.g., handovers' success ratio)
- availability and congestion (e.g., number of transmission time intervals, number of queued users waiting for a resource, congestion ratios, free channels available)

The said antenna KPI data consist of 999,257 observations x 17 features (24 hours cycle), and include more than 40K cell sites. Moreover, the data were anonymised using cryptographic hashing functions from the OpenSSL's libcrypto¹ library and method from the sdcMicro R² package. The sdcMicro package provides a series of probabilistic anonymization methods that depend on a probability mechanism or a random number-generating mechanism, i.e. every time a probabilistic method is used, a different outcome is generated. Since our target variable is a binary label, which corresponds to the notion of "being a hot spot" at a certain day, the data was split into a training (80%) and test set (20%). To this end, we grouped the antennas by their ID, so that each antenna can be either in the train set or in the test set. We sought the splitting that yields as similar percentage of the positive class as possible in the train and test sets.

Finally, for the '*Quality of service in Call Centres*' use case, TID has constructed a real operation dataset from a LATAM (Spanish) country, under the standard CTM format. The anonymised dataset consists of 1.3M transcripts of continuous speech recordings and: 1) does not include personal and company identifiable information (relevant tokens were removed), 2) does not contain speaker information, 3) sentences are switched within the same call, and 4) real timestamps have been obfuscated although the relative order of the calls is kept. From this dataset, a subset consisting of 17K anonymised transcripts was derived, which was further split into 1) train, 2) develop, and 3) test sets. More specifically, all transcripts are labeled with a Customer Satisfaction Index (CSI) score, as indicated by the customer at the end of the call. In addition, the transcripts are being augmented with the output of a sentiment analysis, for both

¹ <https://cran.r-project.org/web/packages/openssl/index.html>

² <https://cran.r-project.org/web/packages/sdcMicro/index.html>

Spanish and English languages. Last, this dataset also serves for benchmarking purposes that is the study of the impact of anonymization in KPIs, e.g., impact on the retrieval of low satisfaction calls but will also be introduced in the Telefonica Hackathon event.

3.2.3 CAIXA datasets

CAIXA generated four different datasets for the I-BiDaaS experimentation and evaluation. It presented three different use cases but one of the use cases was tested with two different datasets (synthetic fabricated dataset and real tokenized dataset).

Synthetic dataset for the '*Analysis of relationships through IP addresses*' use case was the first dataset generated and the one that was selected to test the I-BiDaaS MVP.

The generated dataset provides data on the relationships between customers in order to build part of the social graph of the bank. The data was synthetically generated based on real data coming from a set of restricted tables (relational database), with information related to the customers and their IP address when connecting online. CAIXA and IBM generated the data recipe for the data fabrication using IBM TDF. Through an iterative analysis of obtained results, the rules were improved in order to obtain the fabricated dataset used for testing the MVP, with more than 1 million entries.

The structure of this dataset is the following:

- **FK_NUMPERSONO:** Identifier of the Person. NUMBER
- **PK_ANYOMESDIA:** Date (YYYYMMDD) of the connection of the user. NUMBER
- **IP_TERMINAL:** IP Address of the connection of the user. VARCHAR2
- **FK_COD_OPERACION:** Code of business operation done by the user. VARCHAR2
- **PK_COD_ESTADO_OP:** Code of the status of the operation done by the user. VARCHAR2

This use case was also used for validating the usage of synthetic data as a method to test the performance and adequacy of new technologies before integrating it into CAIXA's premises.

After the generation of the synthetic dataset, a dataset was also created with real tokenized data. The structure of this dataset is the same as the fabricated dataset.

For opening the real data of this use case and the rest of the use cases, CAIXA worked internally on the specification of the types of encryptions that enable the entity to share this data without breaking the privacy of the data and allow a certain level of data analytics over the encrypted data. Indeed, one of the challenges of this approach is to find ways to encrypt the data in a way that loses as less relevant information as possible. CAIXA proposed the '*Advanced analysis of bank transfer payment in financial terminal*' use case in order to first test that new approach in the project, and proposed a tokenized dataset using three different data encryption algorithms (depending on the table field types):

- Format preserving encryption for categorical fields.
- Order preserving encryption for numerical fields.
- A Bloom-filtering encryption process for free text fields.

These types of encryption were also used for the tokenized datasets of the other two use cases ('*Analysis of relationships through IP addresses*' and '*Enhanced control on Online Banking*') and will be further explained in the following subsection.

‘Advanced analysis of bank transfer payment in financial terminal’ use case dataset is a dataset generated collecting most of the relevant and additional contextual information of a bank transfer done by a CAIXA employee in a bank office. It is executed by the employee in the name of a customer that authenticates itself and orders it, by identifying all the relational tables that can contain information of the transfer, the customer, the receiver of the transfer, the office and terminal where it is executed and the employee that proceeded with it. The dataset was used in order to identify anomalies that lead to potentially fraudulent bank transfers or bad practices done in the offices.

The structure of the used dataset is the following:

- **PK_OFICINA:** Office where the transaction has been done. Integer. Format Preserving Encryption.
- **FK_CENTRO_AP:** Office that the employee belongs (generally should be the same as PK_OFICINA). Integer. Format Preserving Encryption.
- **PK_TERMINAL:** Identifier of the workstation of the employee. ID (text). Format Preserving Encryption.
- **IP_TERMINAL_IP:** address of the workstation of the employee. Decimal IPv4 address. Format Preserving Encryption.
- **FK_NUMPERSO_PRINCIPAL:** Identifier of the client (bank transfer sender). Id (text). Format Preserving Encryption.
- **FK_CONTRATO_PPAL_OPE:** Bank account identifier of the sender. 12-digit ID. Format Preserving Encryption.
- **CONTRATO_DESTINO:** Bank account identifier of the receiver. 20-digit or 14-digit ID. Format Preserving Encryption.
- **TIPO_COD_OPERACION_2:** List of the sub-operations to undertake the bank transfer. Comma-separated list of operations identifiers (text). Clear data.
- **PK_TSINSESION:** List of the timestamps of the sub-operations to undertake the bank transfer. Comma-separated list of timestamps. Clear data.
- **PERSONA_LOCALIZADA:** List of the identifiers of the sub-operations, indicating if the client is present or not (S=yes, N=no, A=administrative transactions, -1=Not applicable). Comma-separated list of IDs. Clear data.
- **NIVEL_AUT_REQ:** Level of required authorization. Clear data.
- **FK_EMPLEADO_AUT:** Employee who authorizes the operation (when needed). ID (text). Format Preserving Encryption.
- **FK_IMPORTE_PRINCIPAL:** Transfer amount. Integer. Order Preserving Encryption.
- **IMPORTE_CONOLIDAR_ESTADO_CON:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.
- **IMPORTE_CONOLIDAR_ESTADO_AUT:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.
- **IMPORTE_CONOLIDAR_ESTADO_ANU:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.
- **IMPORTE_VALIDAR_DATOS_NOK:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.
- **IMPORTE_VALIDAR_DATOS_CON:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.
- **IMPORTE_VALIDAR_DATOS_OK:** Other amounts related to the transfer and its sub-operations. Integer. Order Preserving Encryption.

Finally, the dataset for the '*Enhanced control on Online Banking*' use case was generated. This use case focuses on analysing the mobile to mobile bank transfers ordered through online banking (web or application). It focuses on the assessment that the controls applied to user authentication are applied adequately (e.g. second factor authentication) in accordance with PSD2 regulation and depending on the context of the bank transfer.

The structure of this dataset is the following:

- **PK_ANYOMES:** Year and month of the partition, is the one corresponding to the consolidation of the Bizum. NUMBER.
- **PK_ANYOMESDIA:** Year, month and day of the sub-partition, is the one corresponding to the consolidation of the Bizum. NUMBER.
- **PK_IDE_USUCLO_ORIG:** User id number reported from origin. VARCHAR2.
- **PK_IDENUMSESION:** Identifier number of the session in which the transfer is made. VARCHAR2.
- **PK_ORDEN:** Unique number that identifies one of the events of the same transfer for the same use, day and session number. NUMBER.
- **PK_REFERENCIA:** Unique identifier of the transfer, appears in the text lines in the consolidation event. VARCHAR2.
- **IMPORTE_OPERACION:** Amount of money of the transfer. NUMBER.
- **IND_AUTENTICACION_REFORZADA:** Indicator of whether the authentication that requires the transfer is of a strong type or not. NUMBER.
- **TIPO_AUTENTICACION:** Authentication type required to make the transfer. Appears on lines of text in display event. VARCHAR2.
- **AUTENTICACION_ESPERADA:** Total authentications necessary to make the transfer. Appears on lines of text in display event. VARCHAR2.
- **ESTADO_OP_DEC:** Status of the operation that appears in the consolidation of the transfer in the DE_TX_CANAL_LOE table in the PK_COD_ESTADO_OP field. VARCHAR2.
- **ESTADO_OP_DEC_LTX:** Status of the operation that appears in the transfer consolidation in the D the operation has a certain status. VARCHAR2.
- **ESTADO_OP_TRF:** Transfer status in TRF table. VARCHAR2.
- **NUM_PARTICION:** Different partition number that appears in the transfers. VARCHAR2.
- **FK_NUMPERSONO_TIT_LOE:** Identifier number of the person associated with the user's NIF in DATOS_ESPECIALES_NIF_TITULARCLO. NUMBER.
- **COD_TIPO_TERMINAL:** Terminal type code. VARCHAR2.
- **TERMINAL:** Identifier of the Open Line contract when the type of activator is LA and the PAN of the encrypted card when the type of activator is TA. VARCHAR2.
- **EMPRESA:** Company (from CaixaBank group) identifier number. NUMBER.
- **MOTIVO_ANULACION:** Reason why an operation is cancelled. VARCHAR2.
- **HORA_CONFIRMACION:** Time when the transfer is consolidated. TIMESTAMP.
- **FECHA_CARGA:** Data upload date. DATE.

3.2.4 CRF datasets

Initially, CRF gathered structured and unstructured sets of a large amount of heterogeneous data from different sources and different levels. During the preliminary stage, all information has been analysed and CRF interacted with the plant in order to understand in depth the nature of the data.

For the ‘*Maintenance and monitoring of production assets*’ use case, data arrives from sensors mounted on different machines (e.g. linear stages, robots, elevators and so on). The data consists of two different datasets in csv format, named SCADA and MES.

The SCADA dataset contains production, process and control parameters of the daily vehicle production and is structured as follows:

Table 2: Structure of the dataset for the SCADA data

	Id	Value	Unit	Timestamp
Example	667	49.75	mg	23/04/2018

There are over 100 sensors and each one is identified by a specific number (id).

The MES dataset contains specific data associated with the type of vehicle being produced and is structured as follows:

Table 3: Structure of the dataset for MES data

	Date	Time	OP020.Passo[20]	modello_op_020
Format	Date	Hour	Boolean	Number
Example	06/10/2018	09:44:22	0	11

When OP020.Passo [20] changes from 0 to 1, a new vehicle enters into the area with sensors and modello_op_020 indicates the model of the vehicle being processed.

Initially, both types of data were considered, but over time, we faced problems retrieving MES data because of rescheduling activities and changes in the production lines, partially due to the COVID-19 pandemic. Therefore, we decided to utilize only SCADA to obtain thresholds for anomalous measurements for all sensors, also because measurements of sensor identified by the number 141 during 13 days of available MES data showed high variance in measurements even for the same vehicle type. Within the project, by analysing data we found satisfactory information in SCADA data.

For the ‘*Production process of Aluminium die-casting*’ use case, CRF received from the plant different datasets with heterogeneous data (e.g. piston speed in the first and second phase, piston stroke, intensification pressures, temperatures, cooling capacity), quality and operator’s data (e.g., defect manually detect). Due to the complexity of the process, at initial project stages (M9), it was not possible to collect sufficient real data for Big Data analytics, so we extracted the most significant information from all those received and created a synthetic dataset as close as possible to the real data. The aim was to share with the consortium synthetic data to have a flexible and rich dataset to understand the pattern and to develop and test the I-BiDaaS technologies. More specifically, CRF combined into a single file the most significant process parameters that reflect the trend of the real production, in which there are a wide variety and a low veracity of various and heterogeneous information.

The generation of the synthetic dataset has been performed using the IBM’s Data Fabrication Tool (TDF), according to the definition reported in D2.1 [4], and generated a formatted text file, convertible into excel that contains 1 million rows, each of one corresponds to subsequent engine blocks produced.

An excerpt of the structure of the synthetic dataset, detailed in the section 3.9.1.2 ‘*Data description*’ of D2.1 [4] is reported in the table below:

Table 4: Structure of the synthetic dataset for the Production process of Aluminium die-casting

Data type classification		Source/level	Value	Data Type
Example	VMA	Process Parameter	4.95	Number
	ESITO_FILTERED	First operator control result	buono	text
	LP_FILTERED_PROMET	Process parameter	false	Boolean

Synthetic data has been validated with an empirical and analytical technique, as described in section 4.4 ‘Production process of aluminium casting’ of D2.5 [5].

The high-level algorithms, developed in the first part of the project, identified the critical values from the dataset and defined the main parameters that affect the quality of the process. In the meanwhile, we gathered more real anonymised process data. An excerpt of the structure of the real anonymised dataset, containing 187 rows, is reported in the table below with the main parameters identified for the detection of the quality level KPI:

Table 5: Structure of the real anonymised dataset for the Production process of Aluminium die-casting

Data type classification		Source/level	Value	Data Type
Example	JD	Process Parameter	060	Number
	N	Process Parameter	001	Number
	VA1	Process Parameter	0.3156	Number
	VM1	Process Parameter	0.45666161	Number
	Sigma1	Process Parameter	2.38751626	Number
	PM1	Process Parameter	275	Number
	VA2	Process Parameter	1.985034227	Number
	VM2	Process Parameter	2.04331851	Number
	Sigma2	Process Parameter	0.122155264	Number
	PM2	Process Parameter	222	Number
	StatisticDataT2	Process Parameter	0.105999947	Number
	L2	Process Parameter	210.3599854	Number
	StatisticDataT3	Process Parameter	0	Number
	L3	Process Parameter	0	Number
	PM3	Process Parameter	0	Number
	PF	Process Parameter	161	Number
	BH	Process Parameter	56.27001953	Number
	RT	Process Parameter	-0.000997543	Number
	TT	Process Parameter	22.66300011	Number
	Result_1	Process Parameter	1	Number
	Result_2	Process Parameter	2	Number

According to several results, obtained from the CRF internal analyses and during the hackathon, held on M18, we proposed to the plant some strategies to better control data integrity (e.g. a

second level of control for operators' changes in parameters). As a result, two thermal imaging cameras have been installed on all die-casting machines, so that in the penultimate quarter of the second year of the project thermal data, an example of which is shown in Figure 1 and Figure 2, were also retrieved. Therefore, a large dataset of annotated thermal images has been provided in addition to the real anonymised dataset in order to test the complexity of the process with the analytics developed within the project. Several models have been developed to utilize both sensor and thermal images data, as reported in D3.3 [6].



Figure 1: Thermal image in jpg format

37.2	37.1	36.9	37.4	36.6	37.1	36.6	36.6	35.7
37.1	36.7	37.2	36.9	37.4	37.5	36.6	35.4	37.5
37.1	37.1	36.6	36.8	37.7	37.3	36.8	36.5	35.4
37.2	37.3	36.8	37.0	37.5	36.9	36.8	37.2	37.5
36.1	37.3	35.2	37.1	36.8	36.3	35.4	36.9	36.7
35.8	37.3	37.3	37.1	37.2	36.9	36.8	36.2	36.8
36.6	37.3	37.2	36.1	35.9	35.8	36.9	36.0	36.8
36.2	35.8	35.8	37.1	36.2	36.6	36.8	36.0	36.5
37.1	36.8	37.1	36.8	36.8	36.7	35.8	36.6	36.6
35.4	37.0	35.2	36.6	35.7	36.8	36.6	36.4	36.6
36.7	36.7	36.7	37.1	37.3	36.3	36.4	36.4	36.4

Figure 2: Thermal data in csv format

3.2.5 Generic use case datasets

For the generic use cases, three different datasets have been utilised as proof of concept that the provided algorithms work properly.

More specifically,

- A. Coordinates dataset, used for K-Means clustering algorithm. It is a two dimensional dataset containing coordinates (longitude, latitude). A preview is reported in Table 6.

Table 6: Coordinates dataset, used for K-Means clustering algorithm

XXXX	YYYY
5.927870038042057033e+00	-3.760572064416526583e+00
3.667179741171054275e+00	-3.582775947015886420e+00

- B. Coordinates dataset with labels, used for K-Means evaluation algorithm. A labelled csv file with coordinates and the cluster id Number of labels: 10, values [0 – 9]. A preview is reported in Table 7.

Table 7: Coordinates dataset with labels, used for K-Means evaluation algorithm

XXXX	YYYY	Label
5.927870038042057033e+00	-3.760572064416526583e+00	0
3.667179741171054275e+00	-3.582775947015886420e+00	8

- C. ADMM-Lasso dataset. A binary dataset separated in 5 files, containing synthetic values of 10 dimensions.

The provided algorithms of the platform are the following:

- Lasso ADMM: Least Absolute Shrinkage and Selection Operator algorithm for regression analysis, solved in a distributed manner. The LASSO model uses L1 regularization to induce sparsity and prevent overfitting of the model³.
- K-Means - prediction: The objective of K-Means is to group similar data points together in a user-specified number of clusters (K).
- K-Means – evaluation: The objective of this version of K-means is to create a model based on a pre-labelled training dataset that can be used for classification.

3.3 Industrial experiments implementation

In this section, the list of the real experiments carried out using the I-BiDaaS solution is reported. Specifically, for each experiment, the experimental workflow in the operation plan is described, by focusing on its goals and associated questions, as reported in Table 8-10-14-15-16-17-18. All experiments considered the evaluation of usability, operability, robustness, innovation, compliance, privacy awareness and cost of the I-BiDaaS solution and have been used for validating the integrated I-BiDaaS solution in the ‘Co-Develop mode’.

In addition to the eight experiments which relate to specific industrial sectors, two generic use case experiments have also been defined to show how the I-BiDaaS solution can also be applied generally in the ‘Expert mode’ and the ‘Self-Service mode’, considering the perspective of potential non-expert/expert end-users. Furthermore, CAIXA and TID identified the approach to work on cross-sectorial use cases, as explained in subsection 3.3.6.

3.3.1 Experimental workflow

The execution of all the conducted experiments was based on predefined workflows, which allowed better planning and monitoring of the participants. These workflows include a set of action steps that drive the execution of the experiment and correspond to relevant metrics that were captured and used during the evaluation of the results. Although for each experiment, a specific experimental workflow was devised, we can describe the high-level series of steps involved in all experiments as follows:

1. **Project setup:** Definition of the I-BiDaaS experimental setup, internally referred as ‘project’. This includes the name of it, processing mode (batch or streaming data) and input type (single file, directory of file, db source, etc.).
2. **Data selection:** Definition of the dataset to be used for the experiment. This is made by the pilots in the project’s use cases or by the participating end-users in the expert and self-service modes.
3. **Data preparation:** Actions required prior to using the dataset in an experiment. These include actual data collection, legal requirements with respect to privacy and security issues and pre-processing actions depending on the type of data:
 - a. **Real Data** (aggregation, anonymization, encryption).
 - b. **Synthetic Data** (rules generation, fabrication, uploading).
4. **Experiment setup & execution:** Definition of the internal I-BiDaaS representation of an experiment, namely: dataset access establishment, analytics algorithm definition and parameterisation and resources allocation.
5. **Results Visualisation:** Visualisation of the experiment results in various ways according to the nature of data, i.e. from static charts for batch analysis results to real-time interactive graphics for analysis results of constantly incoming streaming data.

³ https://web.stanford.edu/~boyd/papers/pdf/admm_distr_stats.pdf

- 6. Feedback intake:** Assessment of the perceived usefulness of the executed experiment for the participants' internal business operations but also utilisation of results for optimising I-BiDaaS component, e.g. in the cases where data fabrication is used results provide feedback for better fabrication rules to TDF.

The following paragraphs describe how this generic workflow was implemented in each use case and provide an analytical description of the experiment setup, execution and results.

3.3.2 Telecommunication experiments

Telecommunication experiments aim to test the efficiency of the I-BiDaaS solution in the context of improving and optimising current operations. To this end, three operational experiments have been defined, as shown in Tables 9-11.

Within TID, there are different types of users that can benefit from the advanced visualizations and the intelligent dashboards integrated in the I-BiDaaS platform. Considering that this is a heterogeneous group of experts and non-experts, with diverse skills, we can define the following high-level groupings:

- IT & Big Data practitioners: these are employees and third-party consultants with specialised training (e.g., data scientists, software engineers, UX experts) and share a common knowledge on big data analytics.
- Intermediate users: People with basic understanding of data analytics that are used to work with some big data tools, especially for visualisation and big data visual analysis. At best, these users may be able to complete basic data mining tasks using languages like python or R.
- Operators: TID or third-party employees, at different levels in production processes, who need to have access to different cascades and views of the data processing results.
- Non-IT users: People with a very good knowledge of the field and the sector (e.g., product managers, marketing and business units), they could interpret the data but they lack the programming skills or data mining expertise.

With I-BiDaaS, TID tested different Big Data technologies (batch and streaming) in a unified platform to solve the important challenges for the telecommunication sector. In particular, the design and implementation of a complete framework of tools augmented real data platforms with the functionality needed to enable a new, highly diverse and synergistic data ecosystem, in a privacy-preserving manner. Furthermore, advanced visualisation approaches and dashboards allowed to harness the power of multiple heterogeneous sources and big-data analytics. This facilitates the ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - with the primary focus to empower both expert, and non-expert big data practitioners, involved in telecommunication activities.

TID provided three relevant and high-value use cases, as shown in Table 1, which are discussed in the next subsections.

3.3.2.1 Accurate location prediction with high traffic and visibility

This use case aims to analyse the behaviour of local and non-local customers over various periods of time (e.g. holidays), and extract insights on the behavioural patterns of groups of people, enabling them to optimize their value propositions. When users travel around the city they create traffic congestions in network, so it would be useful to forecast immediately next events to anticipate movements at scale and to improve the routing and placement of the telecommunication equipment that is already in place, or to arrange accordingly the new equipment obtained.

For this use case, synthetic and anonymised real data have been used.

The important challenges derived from this use case were to interpolate missing events to recover plausible event trajectories; to minimize processing time with respect to growing data size and to maintain real-time delivery of results.

By selecting the best 1000 models, we obtained a baseline metric of average mean absolute error and it was 1.2565. The baseline model accuracy is decent, although we suspect more accurate model can be made with more data pre-processing (e.g. imputation of missing values). The reader can refer to D3.3 [6] for more details on analytics aspects.

We can predict if a specific antenna will experience a rise of users attached to it, but predictions are not directly linked to specific events or rise of users on the neighbouring antenna. In addition, although the current dataset does not fully support this kind of analysis, this objective remains a possibility for future work, provided that the mobility data are augmented and fused with other datasets, such as weather data, public events and calendar data, other context data, which are necessary to generalize the predictive capabilities of the models.

Because of the data sparsity, predictions are made for 4 hours in front. Using Facebook's time-series tool Prophet, for 1000 best sectors (with the least missing data) predictions on average deviate by 1-2 users compared to the true values.

Figure 3 shows the visualisation of results on the I-BiDaaS platform, utilised in 'Co-Develop mode', with mean number of connected users highlighted in different colours.

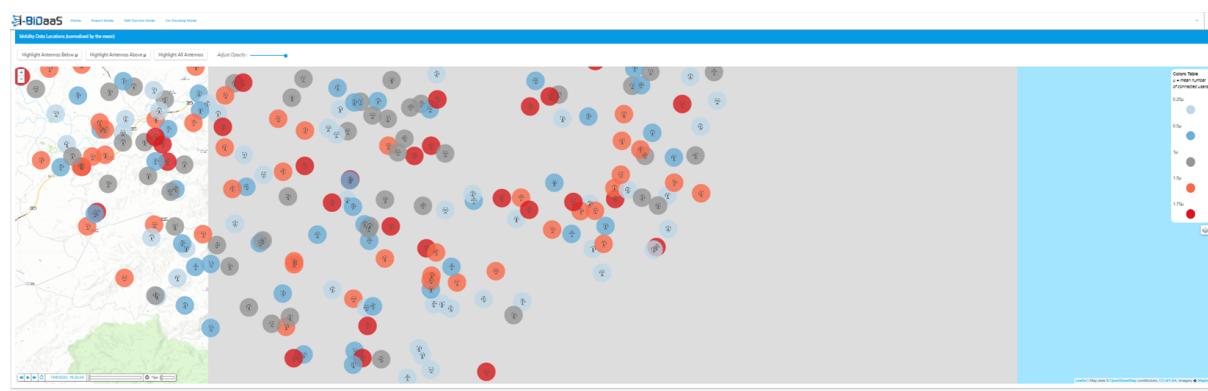


Figure 3: Mobility Data Locations (normalised by the mean)

The following table summarises some of the key points of the use case experimentation:

Table 8: Accurate Location Prediction with High Traffic and Visibility

Experiment #1	Dataset(s):	Preparation status:	Analytics type:
	<i>Anonymised TID mobility data</i>	<i>Generated 100%</i>	<i>batch</i>
	<i>Synthetic TID mobility data</i>	<i>Generated 100%</i>	
Experiment's Goals	To test I-BiDaaS solution efficiency with respect to the prediction of places with high traffic and congestion events in order to optimise their resource distribution.		
Experiment's Questions	<p>Q1. What is the quality of the analytics results?</p> <p>Q1.1 How able is the I-BiDaaS platform to forecast mobile phone user movements at scale?</p> <p>Q2. How efficient is the process of data analytics?</p> <p>Q2.1 Can we predict when new events will cause movements at scale and where will they appear?</p>		

	Q2.2 What is the performance of the predictive models (ML/DL) as a function of time, amount of historical data, and prediction horizon?		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Data preparation <ul style="list-style-type: none"> - Aggregation of antenna KPI data from a major European telecommunications company, covering a city and a large number of cell sites 3 – Data analysis <ul style="list-style-type: none"> - Experiment with various ML models (SVN, RF, XGBoost) and DL models to predict changes in the number of connected mobile phone users per sector 4 – Data visualization		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Data analysts	1-4	4

3.3.2.2 Optimization of Placement of Telecommunication Equipment

This use case aims to optimise the network operations by providing caches and identifying optimal antenna locations, given the provided data from customer usage. The important challenges were analyse streaming data in order to improve the routing and placement of the telecommunication equipment that is available or arrange for new equipment to be obtained; study the spatio-temporal patterns and provide insights on the dynamics of cellular sectors and consider DL models and study their performance as a function of time, amount of historical data, and prediction horizon.

All models – XGBoost, CatBoost and Random Forest showed promising results, while XGBoost stood out with accuracy equal to 0.999, and precision and recall equal to 0.998. Considering the high accuracy and high throughput of the models, the I-BiDaaS solution can help understand network performance, and used as input for network management, planning, and optimization. The reader can refer to D3.3 [6] for more details on analytics aspects.

Currently, near-perfect classification has been achieved, with available data.

Previous analysis on the hotspot prediction task was considered and it brought strong evidence that, for moderate horizons, forecasts can be made even for sectors exhibiting isolated, non-regular behaviour. This work performs forecasts in two situations: daily hot spots and emerging persistent hot spots. We evaluated accuracy as a function of time, prediction horizon, and amount of considered past information. Among others, we observed that the time of the forecast does not significantly affect the results that forecast accuracy reaches a plateau when more than one week of past information is considered, and that tree-based models can outperform the best baseline by 14% on daily hot spots and by 153% on emerging hot spots. In both scenarios, we have seen that the time of the forecast does not introduce a significant variability in the results, and that forecast accuracy reaches a plateau when at least one week of past information is considered. We have also assessed the importance of KPIs in performing such forecasts, showing that this decisively increases for the forecasting of non-regular hot spots, especially for certain usage, congestion, interference, and signalling KPIs.

Figure 4 shows the visualisation of the hotspot prediction on the I-BiDaaS platform, utilised in ‘Co-Develop mode’, for the Optimization of Placement of Telecommunication Equipment.

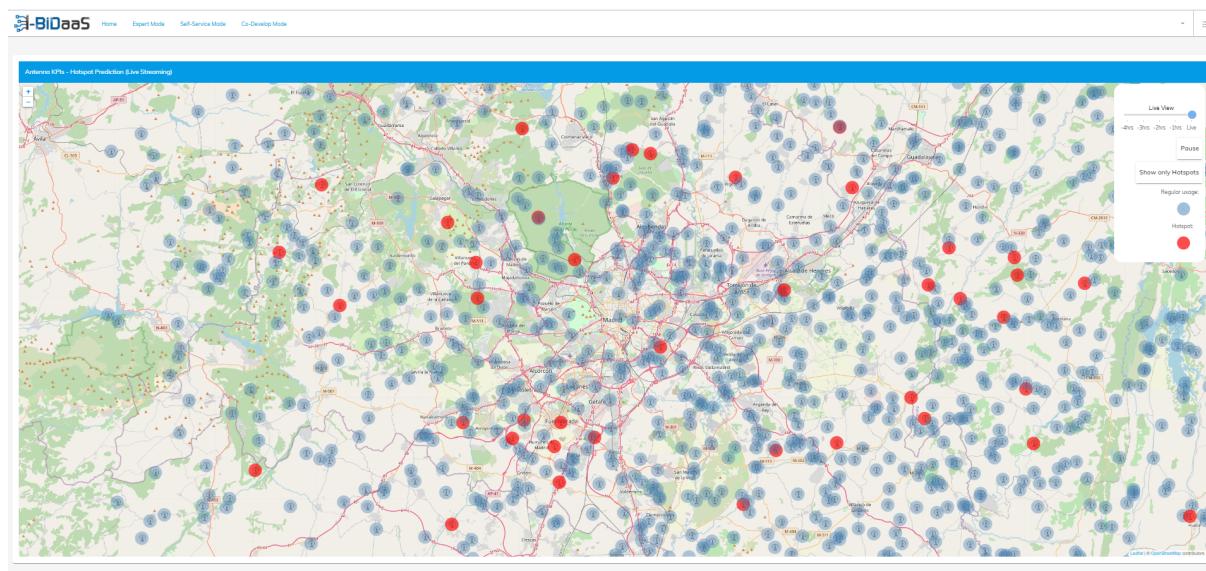


Figure 4: Antenna KPIs - Hotspot Prediction (Live Streaming)

The following table summarises some of the key points of the use case experimentation:

Table 9: Optimization of Placement of Telecommunication Equipment

Experiment #2	Dataset(s): <i>Anonymised TID mobility data</i> <i>Synthetic TID mobility data</i>	Preparation status: <i>Generated 100%</i> <i>Generated 100%</i>	Analytics type: <i>batch</i>
Experiment's Goals	To test I-BiDaaS solution efficiency with respect to the optimization of placement of telecommunication equipment.		
Experiment's Questions	<p><i>Q1. What is the quality of the analytics results?</i></p> <p>Q1.1 How able is the I-BiDaaS platform to support the management of large-scale cellular networks and provide operators with intel on which sectors underperform at any given time?</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.1 Can we timely predict when an antenna will become the next 'hot spot'?</p> <p>Q2.2 What is the performance of the predictive models (ML/DL) as a function of time, amount of historical data, and prediction horizon?</p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Data preparation: - Aggregation of antenna KPI data from a major European telecommunications company, covering a city and a large number of cell sites 3 – Data analysis: - Experiment with various ML models (SVN, RF, XGBoost) and DL models to predict changes in the number of connected mobile phone users per sector 4 – Data visualization		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Data analysts	1-4	4

3.3.2.3 QoS in Call Centres

This use case addresses the challenge of developing speech technologies that transform audio calls into relevant information for the Call Centre, which can be used to assess its performance and/or to screen automatically phone calls. By facilitating the results of the project, TID plans to improve the number of audio calls that can be processed per time unit.

There is a wide variety with respect to the nature of the customer calls: to ask for service and product information, report technical problems, to follow-up with a purchase, to provide feedback, etc. The I-BiDaaS solution allows to quickly get familiar and understand customer's perspective and main interests, and to facilitate a fast response and improve customer service by using big data speech and language analytics. This is achieved by shortening the call durations, waiting time and First Call Resolution (FCR) time by anticipating customer's situation based on previous insights, e.g., using the aggregation of previous analytics by Call Centres or regions.

More specifically, the Business Units (BIs) in TID need to manually inspect a small portion of phone calls, less than 1% of total amount of CC calls per year. The I-BiDaaS solution, using GPU-accelerated text matching, estimates automatically a sentiment score aggregated by call centres/regions and by time window and a list of more relevant words (retrieves top-K frequent words or 2-grams and provides a quick overview of the CC current scenario and operative). In our scenario, we are not interested to compute the sentiment of a certain entity within a transcript, but rather predict the sentiment out of a whole transcript, which will then be further aggregated to predict the sentiment of the whole call centre. As such, the sentiment values of the words that have been found in the text stream of each call centre are accumulated into a single score. This score is actually the sentiment score of a specific call centre for a given time window, and is actually an indicator of whether the overall customer sentiment is positive or negative. In addition, it accounts for the correlation between the sentiment of the call with Customer Satisfaction Index. The reader can refer to D4.3 [7] for more details on analytics aspects.

In Figure 5, an example of the execution of the sentiment analysis tool (English version) in TID server is shown. For a detailed description, see Section 2.5.1 '*Quality of service in call centres*' of D2.6 [8].

```

ow":60,"Epoch":1587628968980}
consumed event ID :7 with TAG : TAG1
Data : {"CallcenterID":"CallCenter1","TopKWords":[{"beautifully":"3.99173699008384"}, {"arbitrary":"3.99173699008384"}, {"annoy":"2.99380274256288"}, {"available":"2.99380274256288"}, {"apathetically":"2.99380274256288"}, {"awkward":"2.99380274256288"}, {"arrogant":"2.99380274256288"}, {"adorer":"2.99380274256288"}, {"abyss":"2.99380274256288"}, {"audibly":"2.99380274256288"}], "SentimentScore": -2, "TimeWindow":3600, "Epoch":1587628968980}
Consumed event ID :8 with TAG : TAG2
Data : {"CallcenterID":"CallCenter1","TopKWords":[{"beautifully":"3.99173699008384"}, {"arbitrary":"3.99173699008384"}, {"annoy":"2.99380274256288"}, {"available":"2.99380274256288"}, {"apathetically":"2.99380274256288"}, {"awkward":"2.99380274256288"}, {"arrogant":"2.99380274256288"}, {"adorer":"2.99380274256288"}, {"abyss":"2.99380274256288"}, {"audibly":"2.99380274256288"}], "SentimentScore": -2, "TimeWindow":86400, "Epoch":1587628968980}

-bash-4.2$ 
SLF4J: Actual binding is of type [ch.qos.logback.classic.util.ContextSelector[23/109]
inder]
[

Sample :[cell[definition=CellDefinition[name='ID' type='Type<Integer>'] value='7'], cell[definition=CellDefinition[name='Word' type='Type<String>'] value='ambitious '], cell[definition=CellDefinition[name='CCID' type='Type<String>'] value='CallCenter1']]]

Triggering GPU
Triggering GPU end
Updating records on UM...

Updating records on UM Done
[Working@es77mx1:~]
```

```

2020-04-23 10:01:06,976 [main] INFO com.tc.l2.logging.TCLogbackLogging - Log file: /home/cle341/Working/store-example04-textual-querying/server-logs/terracotta.server.log
2020-04-23 10:01:06,996 INFO - Available Max Runtime Memory: 1822MB
2020-04-23 10:01:08,053 INFO - ****
2020-04-23 10:01:08,053 INFO - Consistency preference not specified, defaulting to AVAILABILITY mode. *
2020-04-23 10:01:08,054 INFO - This default is deprecated and will be removed in future releases. *
2020-04-23 10:01:08,054 INFO - ****
2020-04-23 10:01:08,054 WARN - It is recommended to keep the total number of servers and external voters to be an odd number
2020-04-23 10:01:08,855 INFO - Becoming State[ ACTIVE-COORDINATOR ]
2020-04-23 10:01:09,157 INFO - Terracotta Server instance has started up as ACTIVE node on 0:0:0:0:0:0:0:9410 successfully, and is now ready for work.

int random = generator.nextInt(upper);
numbers[i]=random;
}
return numbers;
}

public static int addSampleData(DataSetWriterReader<String> writerReader,ArrayList<String>patterns,boolean NormalMode,int start)
int i=0;
int id;
int CallcenterID=1;
int stop = start+10000;
int[] array;
int pat_size=100;
//>100000 >>500
if(NormalMode){
    for(int start;i<stop;i+=500){
        for(j=0;j<patterns.size();j++){
            writerReader.add("callcenter"+(i+1));
            id=CallcenterID;
            Word.newCell(patterns.get(j)+"");
            ID.newCell((i+1));
        }
        CallcenterID++;
        break;
    }
    else{
        for(i=start;i<stop;i+=pat_size)
            array=zeroToUpperBound(patterns.size(),pat_size);
    }
}
```

Figure 5: Example execution in TID server of the sentiment analysis tool.

A human agent can process ~11,520 calls (per year). This manual procedure allows to identify about 2,300 low customer satisfaction audio calls. The I-BiDaaS solution can automatically process ~3.5B calls (per year) in a single GPU. This results in an increase in the number of detected low customer satisfaction audio calls by human agents to 7,000 (200% increase), by pre-processing/filtering the audio calls. This corresponds to a max real-time throughput: 40K transcripts/second.

Given an average call duration of 8.6 minutes, a human agent following a work schedule of 40 hours per week (160 hours per month), could help process up to 11,520 calls per year. This manual process allows to flag ~2,300 low customer satisfaction calls. The I-BiDaaS platform (configuration with 1 GPU) can increase the number of detected low customer satisfaction audio calls by human agents to 7,000 (200% increase), by pre-processing/filtering the audio calls. This corresponds to a max real-time throughput: 40K transcripts/second.

Figure 6 and Figure 7 show the results of the analysis made available to the end-user via the AEGIS's Advanced Visualization tool on the I-BiDaaS platform, utilised in 'Co-Develop mode', that provides an easy tool for end-user inspection and with valuable insights about real-time operations of the CC.

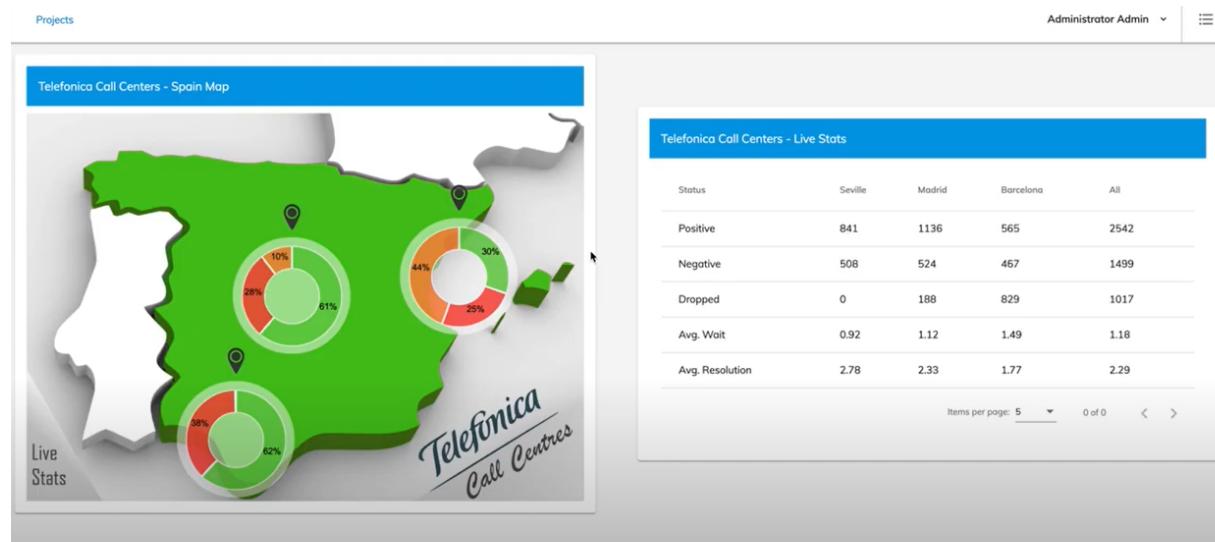


Figure 6: Advanced Visualization Toolkit (AVT) supporting scalable data visualisation.

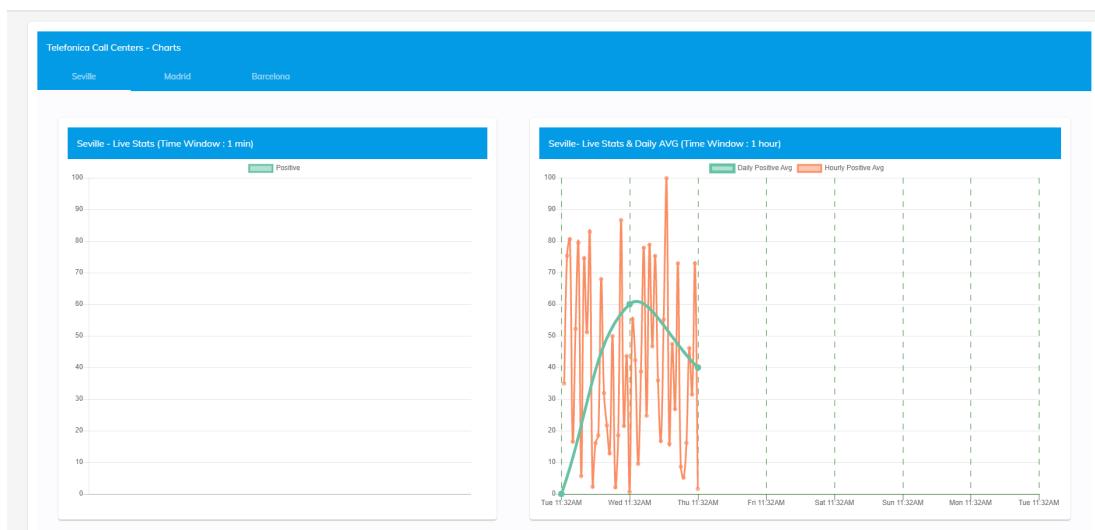


Figure 7: Example of Live Stats & Daily AVG (Time Window: 1 hour)

The cost of the CC is a function of many variables. However, an automatic solution would go a long way to reduce the manual effort and the human resources that need to be allocated for this task. Hence, it will result in a significant reduction of the operational costs.

The following table summarises some of the key points of the use case experimentation:

Table 10: QoS in Call Centres

Experiment #3	Data set(s): <i>Anonymised Call Centre data</i> <i>Synthetic Call Centre data</i>	Preparation status: <i>Generated 100%</i> <i>Generated 100%</i>	Analytics type: <i>streaming</i>
Experiment's Goals	To test I-BiDaaS solution efficiency with respect to the automatic predicting of customer satisfaction.		
Experiment's Questions	<p><i>Q1. What is the quality of the analytics results?</i></p> <p>Q1.1 How able is the I-BiDaaS platform to detect low customer satisfaction audio calls?</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.1 How many low customer satisfaction audio calls can be detected compared to the ones detected by human agents?</p> <p>Q2.2 What is the time reduction obtained?</p> <p>Q2.3 What is the cost reduction obtained?</p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Real data preparation <ul style="list-style-type: none"> - Aggregation of call centre data from multiple sources and generation of data files, e.g., audio files, meta-data, etc. - Execution of ASR model to produce the transcripts - Execution of speaker segmentation to segment the audio files by the different speakers - Merging of the ctm (transcripts) and rttm (speakers) files 3 – Data analysis <ul style="list-style-type: none"> - Apply the NLP model on the input data (merged transcript) to predict the CSI) 4 – Data visualization		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Data analysts	1 – 4	4

3.3.3 Banking experiments

Banking experiments aim to test the efficiency of the I-BiDaaS platform for reducing the costs and time of analysing large datasets whilst preserving data privacy & security. To this end, three operational experiments have been defined, as shown in Tables 14-16.

The usage of big data analytics in the financial sector is every day becoming more and more important and it is gradually being integrated in many departments of CAIXA (security, risks, innovation, etc.). Therefore, CAIXA has a heterogeneous group of experts with different skills and it also relies on several big data analytics experts that give consultancy services. However, the people working with the great amount of data collected from the different sources and channels of CAIXA can be reduced to three groups:

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

Taking that into account, CAIXA 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.

These use cases will be presented in this section in the chronological order in which they were studied in the project.

3.3.3.1 *Analysis of relationships through IP addresses use case*

‘*Analysis of relationships through IP addresses*’ was the first use case, used to test the MVP of I-BiDaaS.

In this use case, CAIXA 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 in their connections to online banking. CAIXA 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 in their revision processes. The goal of this experiment is to validate the use of synthetic data for analysis, if the rules act in the same situations as with the real data and to test the time efficiency of the I-BiDaaS solution.

For this use case, we started using synthetic data, using IBM TDF and the generation process described in D2.5 [5]. The set of rules that build the custom-tailored algorithm to find relationships between two users was identified and the algorithm was programmed in COMPSs by BSC experts. It allowed us to obtain the results on the I-BiDaaS platform, being able to get a first glance visual graphic of the number of relationships, as well as downloading the relationships between customers (Figure 8).

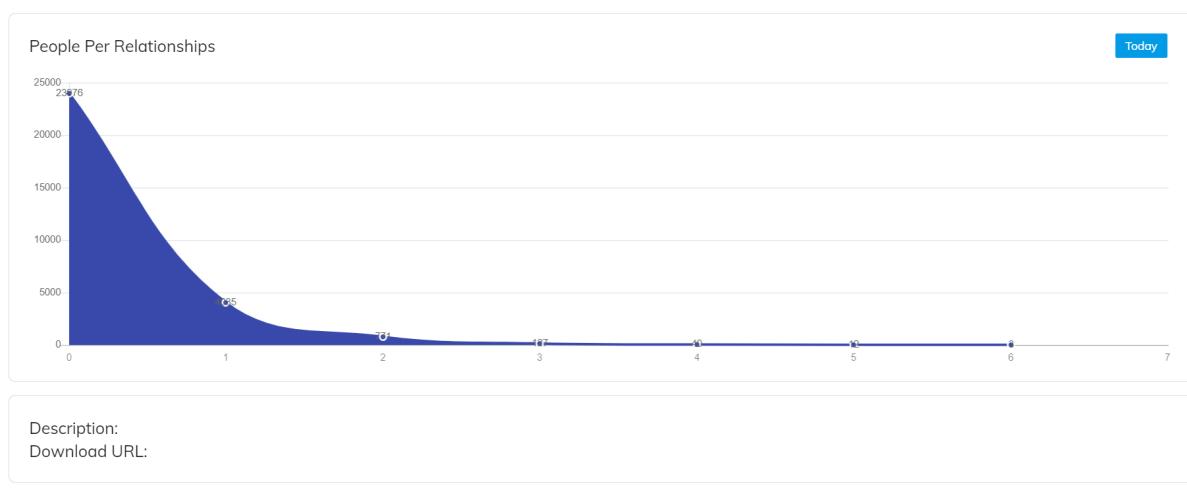


Figure 8: Analysis of relationships through IP address use case visualisation in I-BiDaaS platform

After the generation of relationships, the streaming use case was deployed, using a synthetic dataset of bank transfers between the users and being able to identify those bank transfers in which the sender and the reviewer were already related (Figure 9). In addition, in that use case, the SOC employee can also check the graph of relations of the user, if a more in depth analysis is needed (Figure 10).

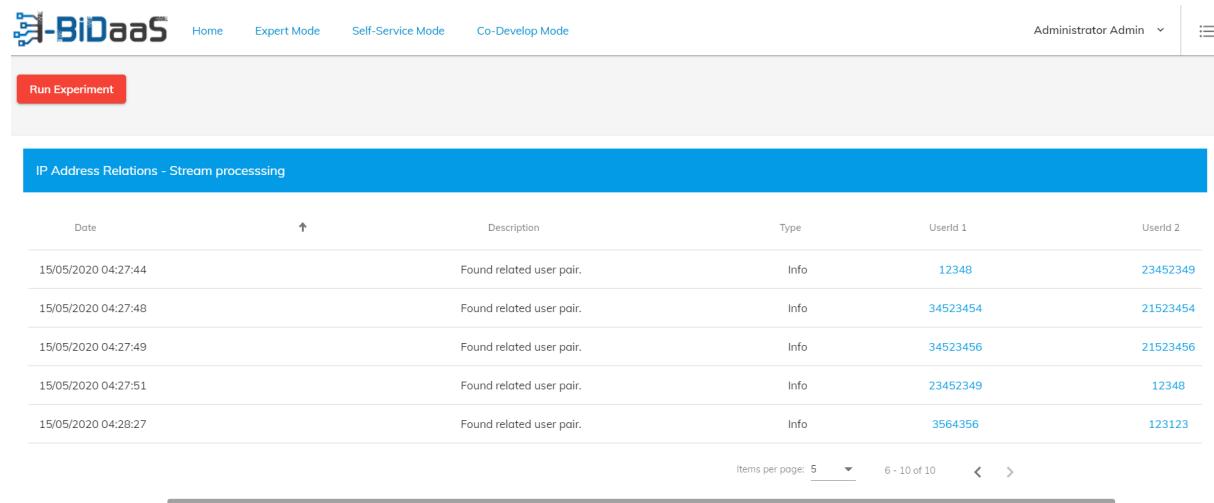


Figure 9: Visualisation of the found relationships between bank transfers' senders and receivers.

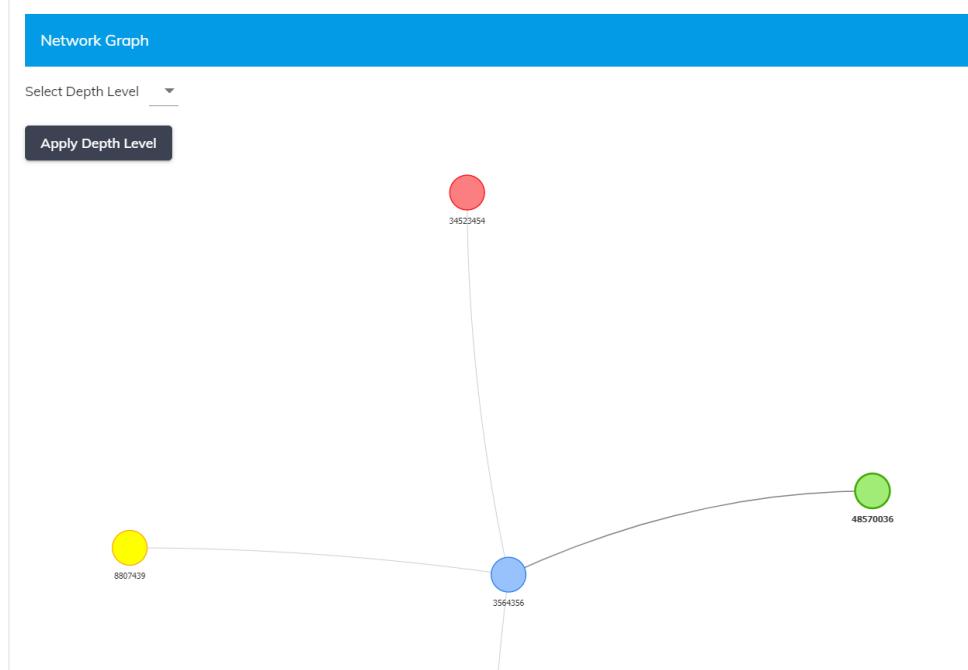


Figure 10: Graph visualisation of the clients' relationships through IP address.

Both use cases were placed in the ‘Co-develop Mode’ of the I-BiDaaS platform due to their personalisation and custom-tailored needs by the I-BiDaaS partners experts.

Moreover, further analysis was performed over the synthetic data generated in order to analyse their quality.

In that process, the dataset was transformed as follows: each user represents a data sample, while each IP address represents a feature. In such a data matrix, the value at position (i,j) represents the number of times user i connects via IP address j . This way, we obtain a data matrix with dimensions 8058x22992. The resulting matrix is very sparse. In order to retain only meaningful data, the next pre-processing step is to drop all the IP addresses that are used by only one user. After dropping all such IP addresses, we are left with 1075 IP addresses, which represents a huge reduction compared to the initial 22992 IP addresses contained in the original dataset.

Next, we filter out the users that are not connected to any of the remaining IP addresses. As it turns out, there are 6049 such users, leaving us with a dataset containing 2009 users and 1075 IP addresses, where each IP address has been connected with at least 2 users.

Clustering

In order to infer relationships between users, we first use clustering algorithms. In particular, we use K-Means [9] and DBSCAN [10]. Additionally, we use t-distributed Stochastic Neighbor Embedding (t-SNE) [11] to visualize the reduced dataset in 2D. The visualization is presented in Figure 11.

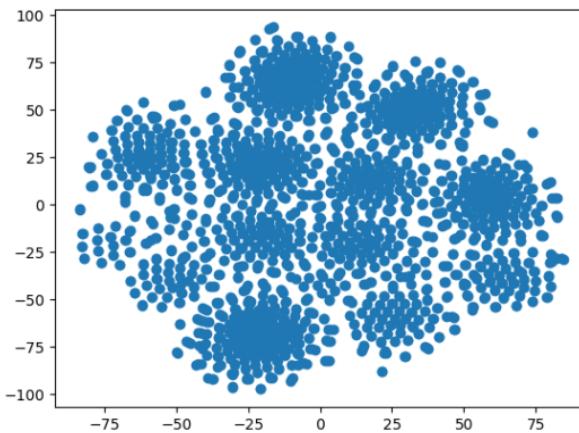


Figure 11: t-SNE projection of the dataset in 2D

Both K-Means and DBSCAN offer some interesting hyper-parameters. In particular, K-Means allows us the flexibility of setting the desired number of clusters. On the other hand, DBSCAN decides on the number of clusters internally. However, it provides the parameters that represent the minimum number of samples in a neighbourhood for a point to be considered a core point. In addition, it provides the maximum distance between two samples for them to be considered in the same neighbourhood. The described K-Means and DBSCAN hyper-parameters are to be set by an end-user based on experimentation and domain knowledge and will be tunable through the I-BiDaaS platform user interface.

Clustering was performed on both the full 2009x1075 dataset, as well as on the t-SNE reduced 2009x2 datasets. We used the silhouette score [12] for evaluating the clustering quality. Roughly speaking, the silhouette score evaluates how well each point fits the cluster it is assigned to versus the next closest cluster. The values range from -1 to 1, with 1 representing a perfect clustering.

Results

All experiments done in both the clustering and graph-based analyses were implemented in Python, on a single computer. Libraries used in a clustering-based analysis are numpy⁴, pandas⁵, scikit-learn⁶ and matplotlib⁷. Numpy and pandas were used for data pre-processing, scikit-learn for clustering analysis, and matplotlib for visualization.

We tested the performances of K-Means and DBSCAN for different values of parameters. In particular, we experimented with the number of desired clusters (K) for K-Means and the parameter for defining the maximum distance between two samples for them to be considered as in the same neighbourhood (eps) for DBSCAN. The parameter representing the minimum number of samples in a neighbourhood for a point to be considered a core point (min samples) for DBSCAN was equal to fixed to 2. This is intuitive since we want to allow the algorithm to find clusters of (at least) 2 people in order to infer relationships.

To start, we applied t-SNE on the transformed dataset in order to visualize the data in 2D. (Any other suitable visualization method to be adopted by I-BiDaaS may be applied.) We set the

⁴ <http://www.numpy.org/>

⁵ <https://pandas.pydata.org/>

⁶ <https://scikit-learn.org/stable/>

⁷ <https://scikit-learn.org/stable/>

parameter perplexity to its default value of 30, while we set the number of iterations to 1000. The algorithm is known to emphasize grouping similar points, so the visual results suggest there is some structure in the data.

For K-Means, we chose 6 different values for K : 13, 600, 700, 800, 900, 1000. The choice is motivated as follows: the visual test, based on t-SNE projection suggests roughly 13 clusters. On the other hand, DBSCAN initially found around 600 clusters. Based on this, we decided to try the small value suggested by t-SNE as well as the higher values suggest by DBSCAN. The results are presented in Table 11.

Table 11: Silhouette score for K-Means clustering

<i>Desired number of clusters K</i>	<i>Silhouette score on full data</i>	<i>Silhouette score on t-SNE data</i>
13	-0.015	0.428
600	0.501	0.578
700	0.601	0.671
800	0.696	0.768
900	0.762	0.856
1000	0.753	0.846

For DBSCAN, we chose 6 different values for eps : 0.1, 0.25, 0.5, 0.75, 1, 1.25. Since 0.5 is the default value, we decided to experiment with values slightly lower and slightly higher than the default. Along with the silhouette scores, we report the number of clusters found by the algorithm for different values of the parameter. The results are presented in Table 12.

Table 12: Silhouette score and number of clusters found for DBSCAN clustering

<i>Maximum distance (eps)</i>	<i>Silhouette score on full data</i>	<i>Silhouette score on t-SNE data</i>	<i>Num of clusters on full data</i>	<i>Num of clusters on t-SNE data</i>
0.1	0.656	-0.689	665	97
0.25	0.656	0.164	665	580
0.5	0.656	0.394	665	706
0.75	0.656	0.507	665	770
1	0.681	0.808	693	882
1.25	0.681	0.815	693	881

Additionally, we evaluated the size of the obtained clusters. A recurring effect is that DBSCAN finds 1 big cluster (e.g., when the number of clusters is 665, DBSCAN finds 1 cluster containing 629 points and when the number of clusters is 693, the big cluster contains 570 points). Interestingly, except from the single big cluster, DBSCAN finds only clusters of size 2, 3, 4 and 5, with clusters of size 2 dominating (e.g. 618 and 643 clusters of size 2 when 665 and 693 total clusters found). As for the K-Means, the algorithm also generates a single big cluster and clusters of size 2, 3, 4, 5, as well as clusters that contain only 1 point. In general, the following pattern emerges: as K grows, the size of the big cluster decreases, but the number of clusters containing only 1 point grows. Once more, the size 2 clusters dominate.

Graph-based analysis

We present here a graph-based solution for relationship analysis among the users. We first describe the data pre-processing steps; subsequently, we describe the methods we use for analysis, and finally, we discuss the obtained results.

The dataset represents the users' online activity made in January and February. In graph-based solution for relationship detection, we did a monthly based analysis, i.e., the results here restrict to the users' activity made in January.

The first step of data pre-processing was to remove the users' activity made in February. The newly produced dataset consists of 71,810 instances. After this, the dataset is transformed in the same way as with the clustering analysis: 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. This way, we obtain a data matrix with dimensions 7947x22680.

Detecting the relationships and generating a graph of relationships

Since, the goal is to establish relationships among the customers, we define an (M, N) – relationship as follows:

Definition 1 ((M, N) – relationship): Two users are said to be in (M, N) – relationship if they have connected at least N times via at least M different IP addresses. Based on this definition, we implement a function (M, N) – relationship. The function takes as input hyper-parameters M and N, and the data matrix. The output of the function is a list of tuples of users that satisfy the condition from definition 1. Next, we use this list to generate a network describing users' connectivity. The vertices in the network represent users. The existence of a link between two users depends on their (M, N) – relationship status, i.e., the link between two users is present if (M, N) – relationship condition is satisfied. Additionally, we implement a function which verifies if two given users are in (M, N) – relationship. The input of this function is a pair of users, the hyper-parameters M and N, and the data matrix; the output is the binary information whether the users are connected or not. On the constructed user relationships graph, we apply the Louvain method for community detection [13]. This is an algorithm designed for detecting communities in networks. It is a simple, efficient, easy to implement and one of the most widely used algorithms for community detection in large networks. We used the python libraries numpy, pandas, networkx⁸, community⁹ and matplotlib¹⁰. Again, Numpy and pandas were used for data pre-processing, networkx and community for community detection via the Louvain method, and finally matplotlib for visualization.

In our experiments, we considered (3, 1) – relationships, i.e., we considered the pairs of users who have connected at least 3 times via at least 1 IP address.

After generating the network, we applied the Louvain method. The algorithm detected 817 communities. Most of the communities were of size 2, i.e., only two connected vertices were forming a community. Table 13 represents different types of components detected via the Louvain algorithm and their frequencies.

⁸ <https://networkx.github.io/>

⁹ <https://python-louvain.readthedocs.io/en/latest/>

¹⁰ <https://matplotlib.org/>

Table 13: Sizes of communities and their frequencies

Size of community (number of users forming a community)	Number of communities of that size
2	604
3	147
4	36
5	15
6	11
7	1
8	3
10	1

Results

The results on the synthetic dataset suggest that both algorithms (K-Means and DBSCAN) manage to group similar users, confirmed by the high silhouette scores obtained by both algorithms. The relatively small cluster sizes (2, 3, 4, 5) suggest that this approach could be meaningful in inferring relationships. Additionally, while the ‘visual test’ after projecting data to the 2D space suggests roughly 13 clusters, the silhouette score confirms that the grouping of data, as well as the possible relationships among them, are much more sophisticated, by assigning higher scores to higher values of K (here K is the desired number of clusters).

In the graph-based modelling approach for relationship inference, we carried out detection of communities in the network of users using the Louvain method for community detection and analysed the obtained results using graph theoretic tools and metrics. The obtained results suggest that the graph-based approach may be suitable for relationships inference. For future work, we will account for the temporal aspect of the IP connections. For example, one can differentiate the relationships obtained on weekdays and weekends and also in working hours and evenings. Such an approach might be able to provide categorization of the obtained relationships (co-workers, friends, spouses, etc.).

The process was repeated with a dataset of real data. After the tokenization (see section 3.3.3.4) of a real dataset of customer connections, the same process was performed and some conclusions were extracted from the comparison between synthetic and real data experiments.

Comparison with respect to synthetic data

The results obtained using K-means clustering on real tokenized data show higher silhouette scores than with the synthetic data, which may suggest that the clusters are in a better agreement with the data than in the synthetic case. Using the synthetic data, only 1 big cluster is obtained, with clusters of size 2 dominating. On the other hand, using the tokenized data, there are 2 big clusters containing most of the points, while the single point clusters dominate. This might offer an explanation for the higher silhouette scores, as single point clusters tend to inflate the metric. Additionally, a large number of single-point clusters on the tokenized real data might point to users that may have specific transaction patterns and might offer an interesting direction for future research and analysis.

The results obtained using DBSCAN on real tokenized data show higher silhouette scores for all, except for the final two values of the parameter eps . Coupled with the drastically decreasing number of clusters found for the final two values of eps (from 681 to 6 clusters found), it can

be hypothesized that, giving a too loose maximal distance between two points to be considered as being in the same cluster causes almost all of the points to be clustered in a small number of clusters. Comparing the results on the real tokenized data with the results on synthetic data using DBSCAN, it can be observed that, with both the real and synthetic data, most of the clusters found are 2-point clusters. However, with real data, two big clusters containing most of the points in the dataset are found, while with the synthetic data, one big cluster was found. Also, increasing the *eps* parameter on the synthetic data results an increase in the number of clusters, as well as the silhouette score. On the other hand, the opposite effect can be observed on the tokenized real data, where increasing the *eps* parameter leads to a decrease in both the number of clusters found and in the silhouette score.

The following table summarises some of the key points of the use case experimentation:

Table 14: Analysis of relationships through IP address

Experiment #6	Dataset(s): <i>1M entries synthetic IP address data; 800k entries tokenized IP address data.</i>	Preparation status: <i>Generated: 100%</i>	Analytics type: <i>batch/streaming</i>
Experiment's Goals	To validate the use of synthetic data for analysis, if the rules act in the same situations as with the real data. To test time efficiency of I-BiDaaS solution.		
Experiment's Questions	<p><i>Q1. Can synthetic data provide the same insights as the real data use case?</i></p> <p>Q1.1 Has the generated data the same structure as the real one? Q1.2 How valid is the model generated with synthetic data with regards to the model of the real data?</p> <p><i>Q2. Is the process of data fabrication more efficient than the process of granting access to real data?</i></p> <p>Q2.1 How much time (mean) is necessary for generating a volume of synthetic data that can provide a valid model? Q2.2 How much is the time reduction that we obtain by generating the synthetic data instead of granting permits to an external provider?</p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Synthetic data preparation <ul style="list-style-type: none"> - generate rules - fabricate synthetic data - upload data set 3 – Real data preparation 4 – Data analysis <ul style="list-style-type: none"> - select algorithm - custom algorithm 5 – Data visualization 6 – Adjust data fabrication rules		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Quality assurance and control managers	3,4	1
	Data analysts	1,2,3	3
	Infrastructure engineers	1, 2, 3, 4	1
	IT security personnel	1, 3	2

3.3.3.2 Advanced Analysis of bank transfer payment in financial terminal

The second CAIXA 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.), employee (i.e. employee id, authorization level, etc.) and bank office (e.g. office id, type of bank office, etc.). All this information is coming from several relational database tables stored in the CAIXA datapool. The meaningful information was extracted and flattered 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 join those ones that relate to the same bank transfer. The heterogeneous nature of the log files, as saved in the CAIXA datapool makes this task even more difficult. There is a total of 969,351,155 events in the log data just for month April 2019. These events are heterogeneous in nature and arise from mixing of disparate operations associated with services provided by the employees in bank offices of different types.

After a laborious table flattening and composition process, a table of 32 fields was obtained and afterwards tokenized according to the encryption schemes described in section 3.3.4. In order to find anomalies through I-BiDaaS, this data was uploaded to the platform through an ‘Expert Mode’ use case, and several algorithms were executed to process the tokenized data, such as K-Means, PCA (Principal Component Analysis) and DBSCAN, as described in section 3.1 of I-BiDaaS D3.3 [6]. Those algorithms were executed using the dislib library with the support of BSC. Although it was studied under an ‘Expert mode’ use case, the complexity of the use case is moderate and it was evaluated by CAIXA that intermediate users could work with it in order to modify parameters of the algorithms and refine the initial anomalies found.

Figure 12 shows the results obtained with the expert mode visualisation of I-BiDaaS.

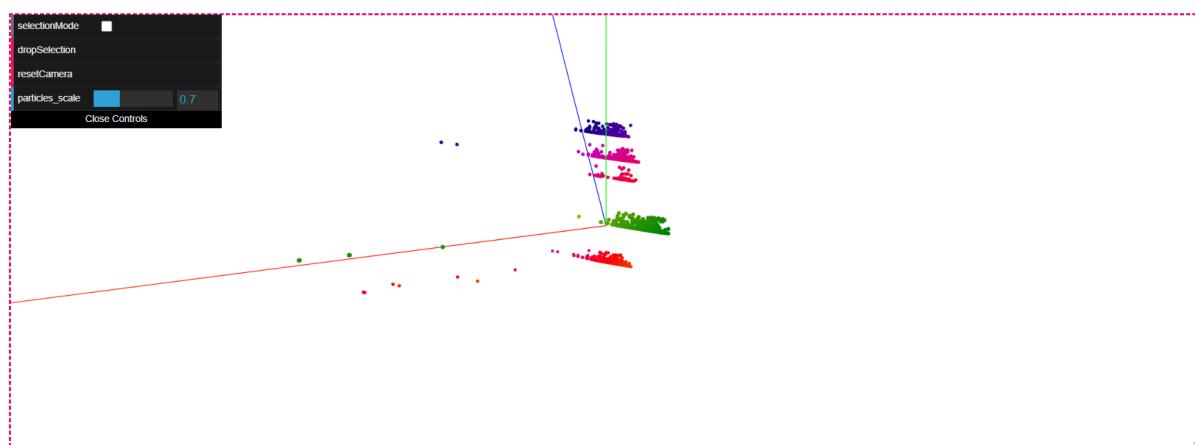


Figure 12: Sample of the ‘Advanced analysis of bank transfer payment in financial terminal’ use case in I-BiDaaS Expert Mode visualisation

The visualisation tool was validated, permitting CAIXA employees to visually identify those transactions that are outliers and select them as anomalies. That allows them to download those anomalies (in several formats such as .csv, .xls, etc.) in order to analyse them on their own or send them to employees of the SOC.

Thanks to this, the tools offered in I-BiDaaS were validated for the full cycle of big data processing, as a self-service for non-IT and intermediate users, while more advanced users are able to customize their big-data analysis. Actually, this gives more flexibility in comparison with competitors. In this sense, we analysed the same dataset with DataRobot, a more mature commercial solution recently acquired by CAIXA. DataRobot provides a benchmark of algorithms that could be applied to an uploaded dataset, and an autopilot option that helps to select the most adequate algorithm. However, the tool is much more focused on supervised learning and it did not provide clear results for this use case, applying any of the unsupervised learning algorithms it provides. Figure 13 shows the results obtained with *Random Forest* algorithm, which was the algorithm that ranked best for the provided dataset. However, they were fuzzy, being much more difficult to identify clear anomalies than in I-BiDaaS.

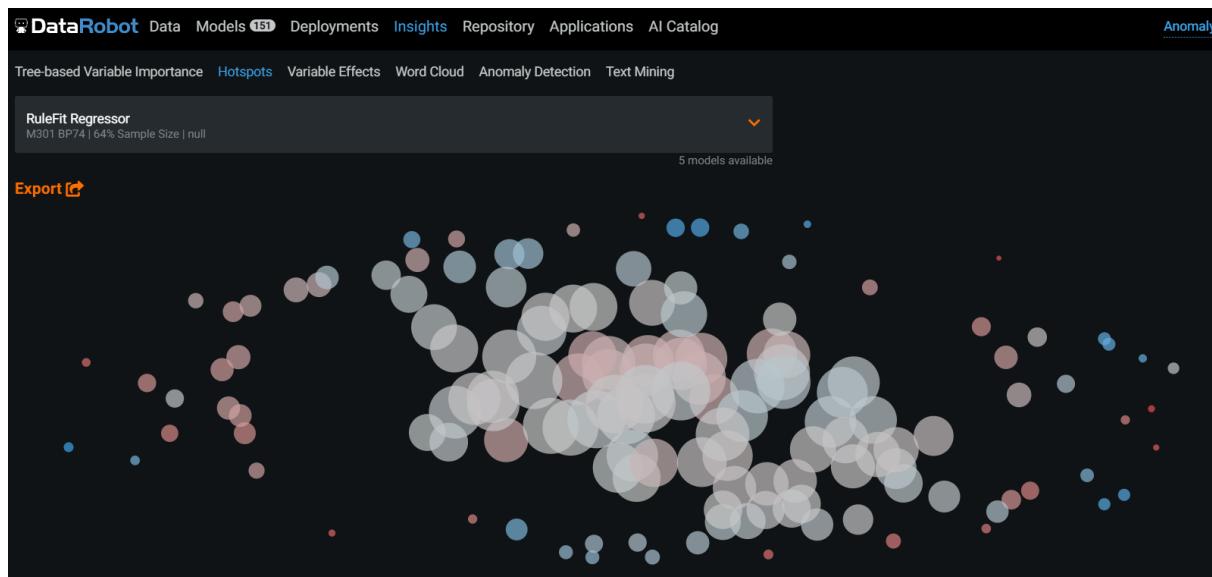


Figure 13: ‘Advanced analysis of bank transfer payment in financial terminal’ use case analysed with DataRobot

As a conclusion of the comparison analysis between the two platforms, on the one hand, we can say that I-BiDaaS provide more flexibility in the definition of your own code, scoring metrics, etc. For example, I-BiDaaS allows to change the scoring function of a specific column, while this feature is fixed in DataRobot. It also allows to provide custom-tailored algorithms or refine them in the case of IT & Big Data expert users.

On the other hand, DataRobot has very a limited number of unsupervised learning models. I-BiDaaS can provide much more detailed results on unsupervised learning use cases based on clustering.

The following table summarises some of the key points of the use case experimentation:

Table 15: Advanced Analysis of bank transfer payment in financial terminal

Experiment #5	Data set(s):	Preparation status:	Analytics type:
	190k entries bank transfer tokenized data (6 months / 100 bank offices)	Generated: 100%	batch
Experiment’s Goals	To test efficiency of I-BiDaaS solution in the context of anomalies detection in the bank transfers from employees.		
Experiment’s Questions	Q1. What is the quality of the analytics results?		

	<p>Q1.1 How able is the I-BiDaaS platform to detect the anomalous bank transfers</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.2 How many potential fraud cases can be solved with I-BiDaaS platform?</p> <p><i>Q3. Does the tokenization/encryption method assure compliance with the current security and privacy regulations?</i></p> <p>Q3.1 Does the tokenization/encryption method ensures the privacy of the data for getting out of the premises of CAIXA without business implications?</p> <p><i>Q4. Which features can I-BiDaaS provide with regards to other data analytics commercial solutions (such as Data Robot)?</i></p>															
Experimental Workflow (based on the generic workflow, to be further refined)	<p>1 – Data selection</p> <p>2 – Data preparation</p> <ul style="list-style-type: none"> - tables identification - data selection - table flattening - table integration - data encryption/tokenization - upload data set <p>3 – Data analysis</p> <ul style="list-style-type: none"> - select algorithm - custom algorithm <p>4 – Data visualization</p>															
Experimental Subjects (participating on any of the steps above)	<table border="1"> <thead> <tr> <th>Role</th><th>Steps involved</th><th>No of participants</th></tr> </thead> <tbody> <tr> <td>Quality assurance and control managers</td><td>3,4</td><td>1</td></tr> <tr> <td>Data analysts</td><td>1,2,3</td><td>3</td></tr> <tr> <td>Infrastructure engineers</td><td>1, 2, 3, 4</td><td>1</td></tr> <tr> <td>IT security personnel</td><td>1, 3</td><td>2</td></tr> </tbody> </table>	Role	Steps involved	No of participants	Quality assurance and control managers	3,4	1	Data analysts	1,2,3	3	Infrastructure engineers	1, 2, 3, 4	1	IT security personnel	1, 3	2
Role	Steps involved	No of participants														
Quality assurance and control managers	3,4	1														
Data analysts	1,2,3	3														
Infrastructure engineers	1, 2, 3, 4	1														
IT security personnel	1, 3	2														

3.3.3.3 Enhanced control of customers to online banking

Finally, in ‘Enhanced control of customers to online banking’ use case, we focused on analyzing the mobile-to-mobile bank transfers ordered through online banking (web and application). It focuses on assessing that the controls applied to authenticate the user are applied adequately (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 is not needed to be encrypted because only a few fields were sensitive. The main objectives of the use case are to identify usage patterns on the mobile-to-mobile bank transfers and enhance the current security identifying the set of transactions in which we should increase the level of authentication. For that reason, we decided to analyze the collected ‘online banking’ dataset and work on non-supervised methods such as clustering of the data. We faced with the need for clustering on a categorical database so that most known algorithms lost efficacy. Initially, an attempt was made to apply a K-Means. K-Means is an unsupervised classification (clustering) algorithm that groups objects into k groups based on their characteristics. Grouping is performed by minimizing the sum of distances between each object and the centroid of its group or cluster. Quadratic distance is often used. Since the vast majority of available variables were not numerical, calculating these distances 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. We transform the variable categories into columns (1, 0), a transformation known as one-hot encoding. This transformation allows 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 [15] 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 minimize 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 *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 iterate 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 minimize inertia (separation between the components of the same cluster).

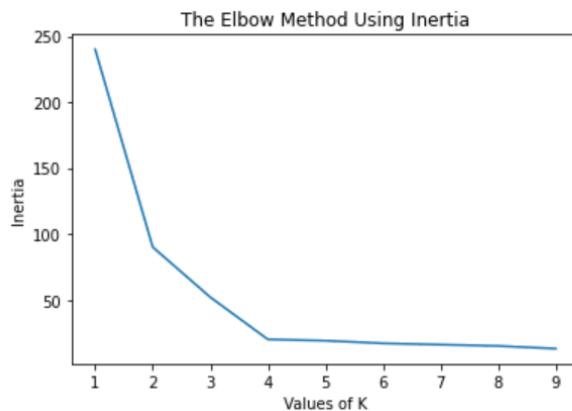


Figure 14: Number of clusters selection for ‘Enhanced Control of customers to Online Banking’

To determine the optimal number of clusters, we have 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.

We worked with BSC in the analysis of this dataset and the clustering of it within the I-BiDaaS platform, being integrated with their support in the “I-BiDaaS expert mode”.

With that support, our ‘Intermediate users’ and ‘Non-IT users’ were able to easily change the number of clusters to run over the dataset and visually analyse the results of it the platform (Figure 15).

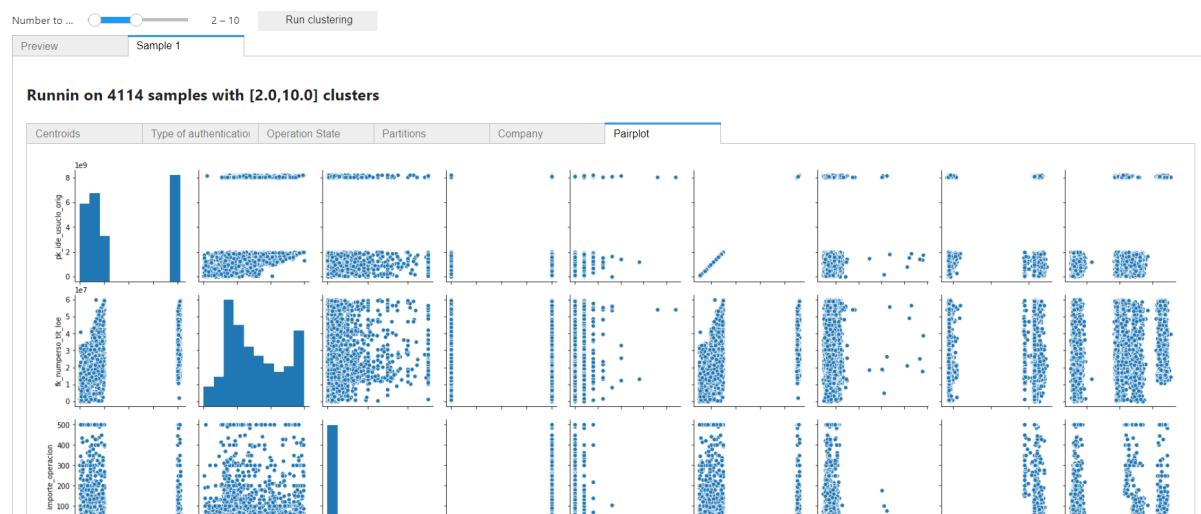


Figure 15: Sample of the ‘Enhanced Control of customers to Online Banking’ use case clustering results in the I-BiDaaS platform

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 patterns of usage of mobile-to-mobile bank transfers and reconsider the way we are selecting the authentication method to proceed with the transfer. Nevertheless, the most important conclusion of the use case was the ability to perform big data clustering analytics in a very agile way, based on existing or custom-tailored clustering algorithms.

The following table summarises some of the key points of the use case experimentation:

Table 16: Enhance control of customers to online banking

Experiment #4	Dataset(s): <i>online banking control tokenized data: 1M+ entries (6 months)</i>	Preparation status: <i>Generated: 100%</i>	Analytics type: <i>Batch/streaming</i>
Experiment’s Goals	To assess that the controls applied to authenticate the user are applied adequately (e.g., second-factor authentication) on mobile-to-mobile bank transfers in online banking.		
Experiment’s Questions	<p><i>Q1. What is the quality of the analytics results?</i></p> <p>Q1.1 How able is the I-BiDaaS platform to cluster the dataset into meaningful datasets?</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.2 How easy was to run the clustering and identify potential errors in the customer authentication?</p> <p><i>Q3. Does the tokenization/encryption method assure compliance with the current security and privacy regulations?</i></p> <p><i>Q4. Which features can I-BiDaaS provide with regards to other data analytics commercial solutions (such as Data Robot)?</i></p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Data preparation <ul style="list-style-type: none"> - tables identification - data selection - table flattening 		

	<ul style="list-style-type: none"> - table integration - data encryption/tokenization - upload data set <p>3 – Data analysis</p> <ul style="list-style-type: none"> - select algorithm - custom algorithm <p>4 – Data visualization</p>		
Experimental Subjects <i>(participating on any of the steps above)</i>	Role	Steps involved	No of participants
	Quality assurance and control managers	4,5	1
	Data analysts	1,2,3, 4	3
	Infrastructure engineers	1, 2, 3, 4, 5	2
	IT security personnel	1, 3	1

3.3.3.4 Data Encryption

During the I-BiDaaS project life-time, CaixaBank changed its approach with regards to how to extract sensitive data and allow big data analytics outside its premises, thus breaking inter- and intra-sectorial data-silos, and support data sharing, exchange, and interoperability. At first, in the project definition, it was planned to evaluate and validate only synthetic data generation and the usage of this synthetic data with I-BiDaaS tools. However, after the first use case, we realised that relying only on synthetic data was a limitation for extracting new insights from the data. Therefore, CAIXA moved into a more open position, starting to evaluate ways to share real data. In that sense, to facilitate sharing of such data to a third party (i.e. uploading dataset to any external cloud) requires a cryptographically secure encryption process without degrading the quality of data. Several data encryption experiments were undertaken during the project and used for the data tokenization in the use cases.

In this section, we provide a proof of principle demonstration of the encryption schemes that were used in the project for encrypting financial data.

Format-Preserving Encryption

Our principal aim is to encrypt sensitive data with the constraint that encrypted data closely follow the real data.

For example, encryption of a string with 4-digits will give back another string of 4-digits. For this, we use an encryption scheme known as **Format-Preserving Encryption**: the goal of a Format-Preserving Encryption scheme is to securely encrypt the data while preserving its original format.

To carry out such a form of encryption, we specifically used the algorithm **Feistel-based encryption (FFX)¹¹**. The construction we have used is based on hash-based pseudo random function **HMAC**¹².

Order-preserving Encryption

Order-preserving encryption (OPE) allows to compare cipher text values in order to learn the corresponding relation between the underlying plaintexts. By definition, order-preserving

¹¹ <https://csrc.nist.gov/csrc/media/projects/block-cipher-techniques/documents/bcm/proposed-modes/ffx/ffx-spec.pdf>. Method 2 was used in the algorithm.

¹² <https://en.wikipedia.org/wiki/HMAC>.

encryption methods are less secure than conventional encryption algorithms for the same data sizes, because the former leak ordering information of the plaintext values.

CAIXA stores several numeric data that must be secured before sharing. This method is quite useful when you want to apply external algorithms which use ordered data. A typical example is when you want to compare the Age of two clients. In these cases, we would have:

$$\text{Age(A)} < \text{Age(B)} \rightarrow \text{Encrypt}(\text{Age(A)}) < \text{Encrypt}(\text{Age(B)})$$

The implementation of this algorithm is leveraged to open source libraries. See **pyope package**¹³ for more information.

Privacy-preserving encryption of text using Bloom filters

In the financial data owned by CAIXA, one kind of sensitive information consists of free texts like the surnames, street names etc. Such information can be of importance for establishing relations between clients. Using encryption schemes, such as described in Section 3.3.3.4, one can create complications in establishing relations. The main reason is that any mistake in spelling (or different way of writing, for example: *L'hospitalet* and *Lhospitalet*) will create encrypted texts totally different from the original.

To address such an issue, we use a privacy-preserving scheme used for record-linkage [14] by employing cryptographic Bloom filters. We will describe in detail the process of record-linkage in the following subsections. We just comment here on one important property: Bloom-filter based cryptographic schemes are non-reversible. This is to say that, just knowing the private keys and encrypted data, one cannot go back to the decrypted original data.

The encryption is carried out in the following steps:

Splitting text in n-grams

In the field of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech¹⁴. For example, an example of 2-gram splitting of the word **L'hospitalet** with a slide/window of step 1 is:

2-gram(L'hospitalet) \Rightarrow [L', 'h, ho, os, sp, pi, it, ta, al, le, et]
--

Ngram

To create a n-gram from a text

Can be found in the folder:
encriptar_tabla_aplanada/src_encrypt/encryption/encrypt.py

Creating Bloom-filter atoms

Atoms denote Bloom filters that are generated from only one bigram. For example, a Bloom-filter atom looks like:

Bloom-filter atom (L')



0	0	1	0	0	1	0	1	0	0	0	1	1	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

¹³ <https://pypi.org/project/pyope/>

¹⁴ <https://en.wikipedia.org/wiki/N-gram>

A Bloom-filter atoms uses cryptographic hash functions followed by a mapping to a bit-array. We have used two set of parameters depending on the type of text one need to encrypt. The texts which are not susceptible to frequency-based attacks¹⁵ (i.e. street name), we follow the procedure outlined in the article ‘*Privacy-preserving record linkage using Bloom filters*’ [14] with the following parameters:

- Two independent hash functions (sha224, sha256).
- Number of dependent hash functions.
- Length of bit array.

For texts susceptible to frequency-based attacks (surnames)¹⁶, we increase the number of independent hash functions along with the bit array size as per the suggestions¹⁷.

Cryptographic Bloom filter and comparison

The cryptographic coding of a text is carried out by collecting all bit arrays created from the application of Bloom-filter atoms on each element of the n-gram of the text. Then the collection of bit arrays are joined by Boolean OR operations. Such a cryptographic procedure can be used to find the closeness of two texts by comparing similarity measures. An example of Bloom-filter representations is shown pictorially in Figure 16, taken from the aforementioned article [14], of two similar texts (A: SMITH, B: SMYTH).

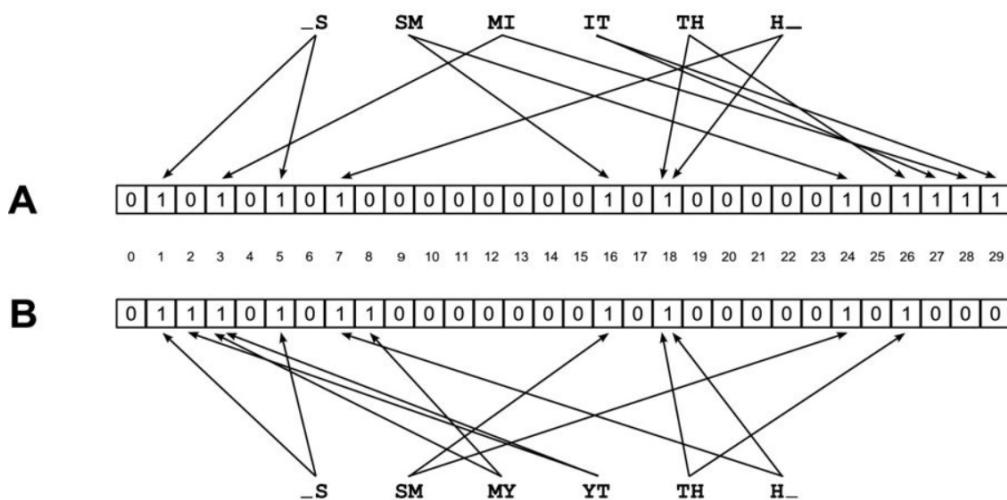


Figure 16: Bloom-filtered arrays resulted from two similar texts.

The similarity measure we use is called **Dice coefficient** and is given by

$$D(A, B) = \frac{2h}{a + b}$$

where h is the number of bit positions set to 1 for both A and B, a and b are respectively the number of bit position set to 1 in A and B.

Figure 17 shows how a Bloom filter can be used to single out similar street names.

¹⁵ Frequency-based attacks details: https://en.wikipedia.org/wiki/Frequency_analysis

¹⁶ <http://openaccess.city.ac.uk/14304/>

¹⁷ <http://www.numpy.org/>

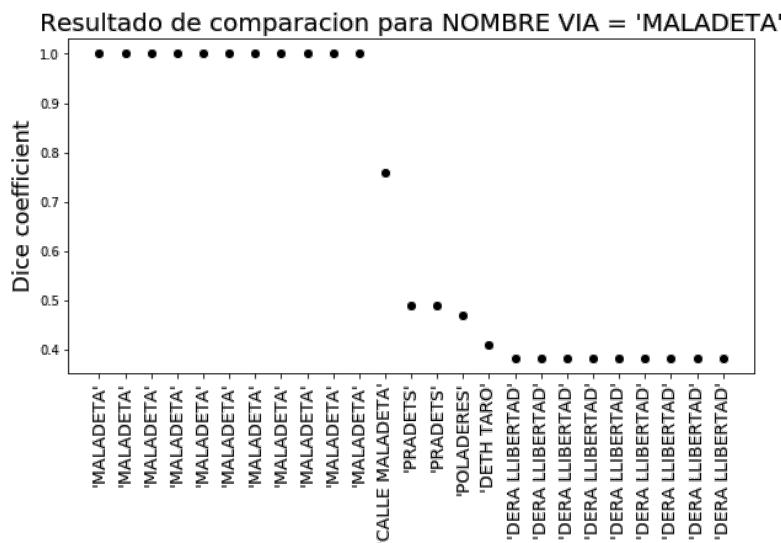


Figure 17: Result of street name similarity after being encrypted with Bloom-filters.

Other evaluated encryption methodologies

Fully Homomorphic Encryption (FHE) was also considered as a potential encryption scheme for some data fields. FHE is a cryptographic technique that allows to perform operations on encrypted data that are equivalent to directly manipulating the plaintext. Performing analytics over encrypted data has an intrinsic trade-off: Accuracy-Security-Performance. Accuracy is measured against the accuracy of comparable plaintext analytics; Security is measured in terms of the ability to deduce information about the private encrypted data; Performance is measured against the time and storage performance of comparable plaintext analytics. For complex tasks, in most cases, at least one of these elements is sacrificed for the others.

All existing FHE schemes have the property that the encrypted data contains noise, and this noise increases when this data is manipulated. When performing long computations, this noise needs to be cleaned every once in a while. This can be done in two ways. One way is to interact with the client (the owner of the data who encrypted the data in the first place) as follows: Every time a ciphertext accumulates too much noise, it is sent back to the client, where it is decrypted, encrypted again, and returned. Decrypting cleans the noise and encrypting again creates a fresh ciphertext with minimal noise. Another way that is completely non-interactive is to use an operation called Bootstrapping, which cleans the noise. This operation is computationally expensive and currently not available in most FHE schemes implementations.

The main experiment performed focused on training a complex Neural Network (NN) under FHE. As the complexity of the NN grows, the time and storage overheads, introduced by FHE, become quite big. Overcoming this challenge, while maintaining acceptable scores in all three metrics: Accuracy-Security-Performance was the challenge that we set out to solve in the experiment. We worked simultaneously on the simplification of the NN architecture and on HE optimization to improve performance while maintaining industry acceptable security levels and minimizing accuracy degradation. We were able to train a complex NN within an 8-hour timeframe with an accuracy degradation, which is linearly dependent on the number of CPUs used. Currently, for 6 output classes and 24 CPUs, the accuracy drop is about 20%, and for 2 output classes, there is hardly even an accuracy drop. Due to the parallelization work that was done, this figure can be improved by increasing the number of CPUs.

However, this experiment focused on this very specific experiment and we concluded that currently, it was not applicable generally for the financial data encryption.

3.3.4 Manufacturing experiments

The main objective of the manufacturing experiments is to demonstrate the ability to exploit Big Data in order to take advantage of the real-time shop-floor data to apply sophisticated statistical assessments. Tables 17-18 present the definition of the associated experiments.

Manufacturing 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 do 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.

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 implement in 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. Together with the experimental subjects who participated in the experimental workflow, the final end-users for the manufacturing sector can be grouped into three main groups:

- Manufacturers: People who have the relevant experience and current practices to innovate and improve, and offering the opportunity to validate and demonstrate the project, its approach and results across real contexts.
- Intermediate users: People involved in data collection, data security, manual analysis, operational flows and required functionalities by investigating I-BiDaaS solution in order to innovate the production management processes.
- Operators: People employed at different levels in production processes, who need to have the data processing results really useful. This is achieved, for example, through advanced data visualization methods that provide the insights, value, and operational knowledge extracted from data available. This system allows the operator to understand the meaning and relationships of the analysed data, through graph representations of the algorithms developed by the consortium.

As an industrial end-user, CRF identified the necessary requirements to develop analytics on the retrieved data from real industrial environment. For both use cases, confidentiality is very important to protect information from being accessed by external parties. So data have been anonymised before shared. Furthermore, the lack of time to extract and analyse data due to the fast rhythms of production and fast internal changes due to rescheduling production quantities and component variations required data cleaning in terms of identification of incomplete, inaccurate and irrelevant parts of the data and data analyses with advanced visualisation tools to better empower manufacturers decision. All analyses were carried out by I-BiDaaS experts, as detailed in the corresponding activities developed within the entire project and reported in all technical deliverables. In the next two sections, we describe the main outcomes for both use cases and explain how we can use I-BiDaaS solution to develop a methodology to implement the use cases in real scenarios for quality and process improvements and Predictive Maintenance.

3.3.4.1 Maintenance and Monitoring of production assets

This use case has been selected to use the data to optimise a real industrial process and to set a predictive maintenance procedure in order to prevent faults before they happen by doing maintenance at the right time (not too late or too early, to avoid inefficiencies). Different types of sensors are installed on the production line and acquire different data information (e.g. acceleration, velocity, pressure, temperature and so on). All of the sensors record their

perception of the surroundings, uploading and transfer this information to a server that manages the data. For example, accelerometers are used for measuring vibration and shock on machines and basically anything that moves. Therefore, the monitoring of vibrations is important to check the status of a machine and the analysis of the trend of vibrations over time allows to predict the onset of deterioration and to intervene in time before the failure. The continuous and periodic control of the service conditions of a machine is known as Predictive Maintenance. The goal of this experiment is to test the I-BiDaaS platform, using different methods adapted to different users (expert/ non-expert) across silos: different companies, departments and competences are involved.

Before analysing, data have been transformed into separate time series - one per sensor - in order to monitor the separate time series for each sensor any day. As described in D3.3 [6], I-BiDaaS analysts carried out an outlier detection analysis on each sensor separately. Subsequently, they compared the time stamps of the detected anomalous measurements across the results for different sensors. The analysis did not require any parallelization, everything was done on a single GPU (NVIDIA RTX2070). The outlier detection analysis was performed using a modified interquartile range (IQR) test. It was established that almost all sensors have different days with anomalous measurements, and almost all of them were common to different sensors (more than 90% on average). Two more similar tests were performed, where Q1 was calculated as the 10th (5th) quantile and Q3 was calculated as 90th (95th) quantile. The most informative results were obtained for Q1 = 5th, and Q3 = 95th percentile.

After implementing these results, the efficiency and accuracy of the I-BiDaaS model with respect to internal CRF analyses allowed to quickly visualise the results on the I-BiDaaS platform, in the ‘Co-Develop Mode’, being able to get a visual graphic of the anomalous measurements for the selected year, month, day and sensor, as shown in Figure 18 and Figure 19, by giving us the possibility of developing a methodology to intervene with specific actions.

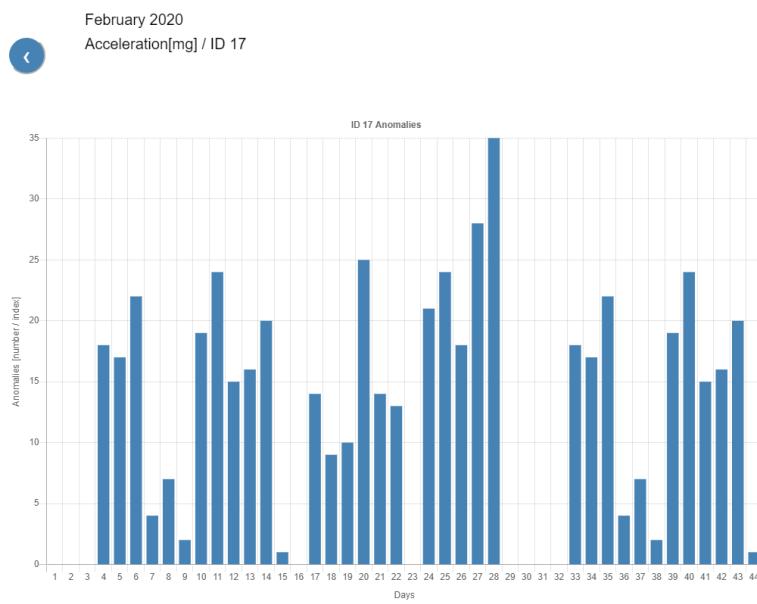


Figure 18: Anomalous number per day for the selected sensor and month

By pressing on any bar, in Figure 18, it is possible to visualise the anomalous values for the selected day and match them to try to understand what happened.

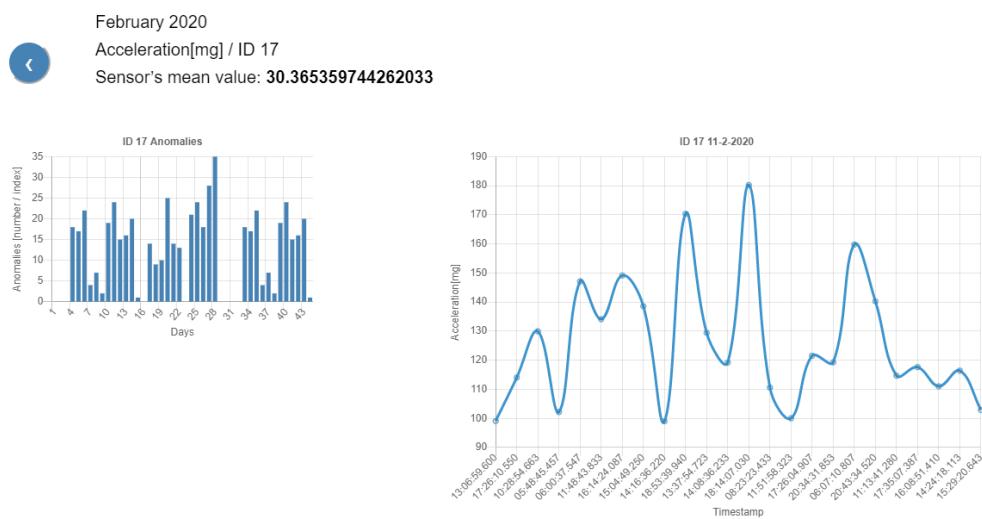


Figure 19: Sensor's mean value and anomalous trend for the selected sensor and day

Based on these results obtained with batch analytics, the consortium worked on the creation of a structured foundational database to be easily utilised to check outliers for the continuous and periodic control of the service conditions. Figures 20-22 show that it is possible to select year, month and day, the category of the sensor and the specific number of sensor to visualise the number of anomalies per day and their trend, as reported in Figures 18-19.

Monitoring of production assets - CRF/FCA

Choose a month to proceed with the analysis

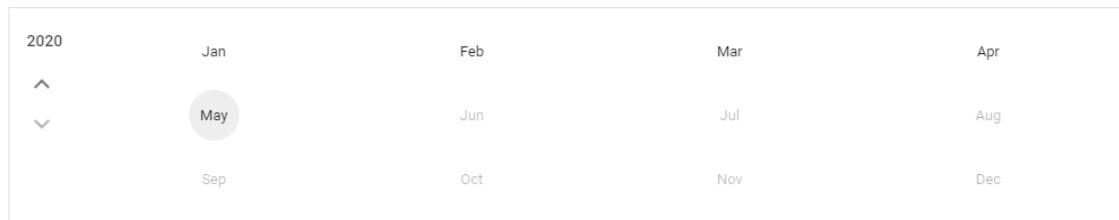


Figure 20: Sensor calendar

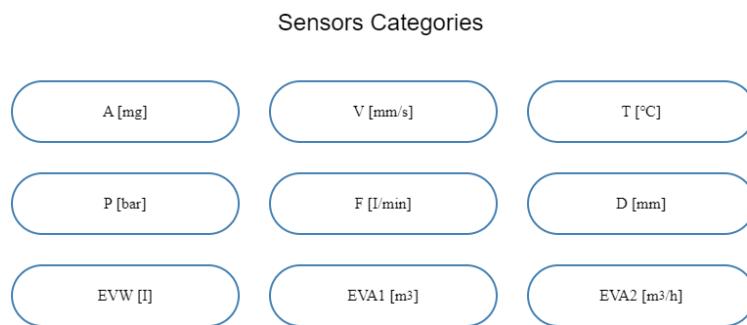
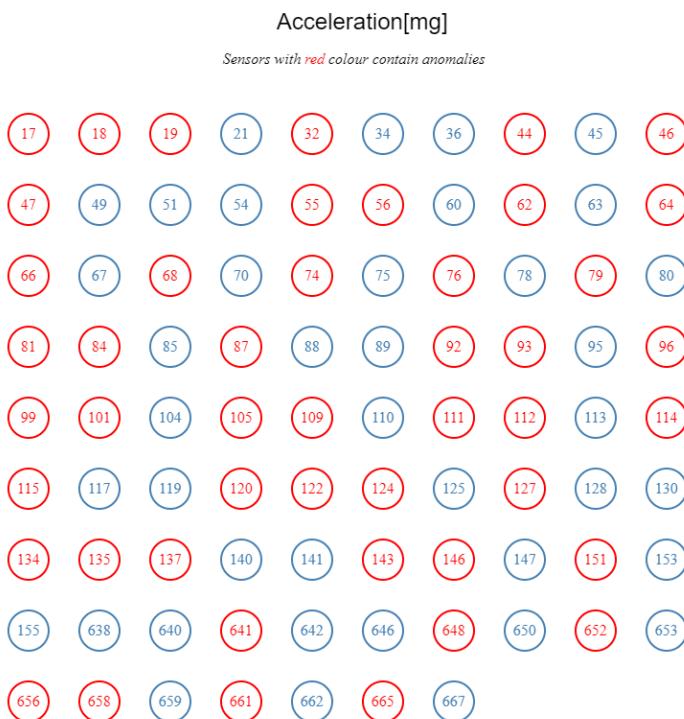


Figure 21: Categories of sensors

**Figure 22: Sensors of acceleration**

All this leads to the prediction of unnecessary actions and improves the efficiency of manufacturing plants, by reducing production losses and achieving greater competitiveness of the company by an increase of 0.05 % the current Overall Equipment Effectiveness (OEE) and a decrease of 50 % in maintenance costs.

The following table summarises some of the key points of the use case experimentation:

Table 17: Maintenance and monitoring of production assets

Experiment #7	Dataset(s): <i>Real MES</i> <i>Real SCADA</i>	Preparation status: <i>Generated 100 %</i> <i>Generated 100%</i>	Analytics type: <i>Batch</i>
Experiment's Goals	To test efficiency of I-BiDaaS solution in the context of anticipation of maintenance events (alarm).		
Experiment's Questions	<p><i>Q1. What is the quality of the analytics results?</i></p> <p>Q1.1 What is the accuracy of new models with respect to internal CRF models in use (geographical representation of the process)?</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.1 How efficient is the performance of the analytics application (algorithm)?</p> <p>Q2.2 How efficient is the visualisation of the analytics solution to allow the workers a quick intervention with specific actions?</p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Real data preparation 3 – Data analysis <ul style="list-style-type: none"> - select algorithm - custom algorithm 4 – Data visualisation		

Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Quality assurance and control managers	1, 4	1
	Data analysts	1- 4	2
	Financial administrators	4	1
	Infrastructure engineers	1	1
	IT security personnel	1	1

3.3.4.2 Production Process of Aluminium die-casting

This use case aims to improve the quality of the production process of the engine blocks. During the die-casting process, molten aluminium is injected into a die cavity where it solidifies quickly. The process is complex and it is important to not only carefully design parameters and temperatures but also to control them because they have a direct impact on the quality of the casting. Big Data analysis aims to improve the quality of the process, with the aim of finding the most significant parameters to monitor and control. The goal of this experiment is to test the efficiency of the I-BiDaaS solution in the context of correlating defects with the production process parameters and resetting these to prevent repairs and reprocessing of the engine blocks.

Firstly, to allow the classification of the engine blocks according to their control class, Random Forest has been used to assess the feature importance and possibly point to the most important parameters in the process that determine the outcome of the classification. The first analysis showed that this use case corresponds to an imbalanced problem, i.e. there are more samples belonging to class ‘a’ than to class ‘b’, so the basic random forest algorithm is susceptible to discriminate and favour the larger class. To avoid this scenario, by associating different weights (or rewards) for different classes in the objective function, weighted Random Forest has been used, an extension of the basic random forest, by associating different weights (or rewards) for different classes in the objective function. Treating the problem in the use case as a binary classification problem, with the aim of identifying scrap and proper engines, a binary classification algorithm has been applied. Furthermore, the newly implemented distributed alternating directions method of multipliers (ADMM) algorithm has been applied and has been used to perform the binary classification on the given dataset. Since the data contains a large number of process parameters, and because of their proven performance in practice, the Deep Neural Network framework has been chosen. Deep Neural Networks contain a large number of layers that enable them to learn relationships between the input data and the target values, whether that relationship is linear or highly non-linear. Considering that the part of the data for this use case consisted of thermal images of engines, convolutional neural networks, and DenseNet-201 in particular, have been used in image classification. Further details on modelling and analytical approach for this use case can be found in D3.2 [18] and D3.3 [6].

A visualization approach using t-Distributed Stochastic Neighbor Embedding (t-SNE [11]) is also used to visualize the data in 2D and see whether there is any structure emerging (see D3.2 [18]).

To develop a methodology to improve the quality of the process, we can quickly visualise the results of the analyses, performed by I-BiDaaS experts in the I-BiDaaS platform. In the following figures, the self-service solution, developed by I-BiDaaS, is explained step by step.

A dynamic diagram shows the incoming streaming data in real-time, as well as aggregations of them that are constantly updated, after pressing the top-left button ‘Run Experiment’.

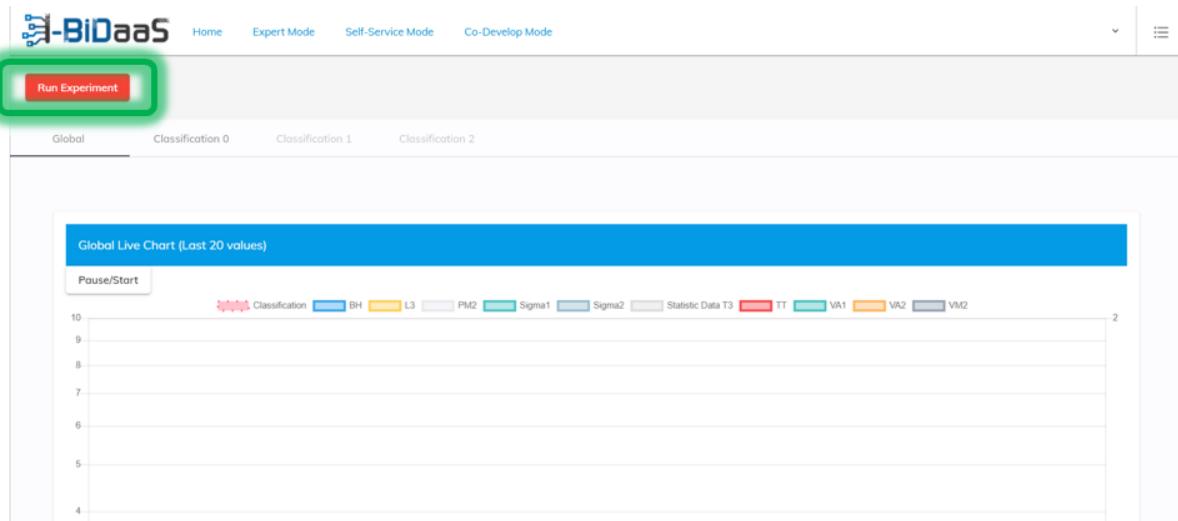


Figure 23: Dynamic diagram for the ‘Production process of Aluminium die-casting’ use case

We can visualise the trend of parameters and quickly check if there is an anomaly compared to the set values of parameters. If we see the anomaly, we can press on the button ‘Pause/Start’ and quickly visualise data trends in the selected range.

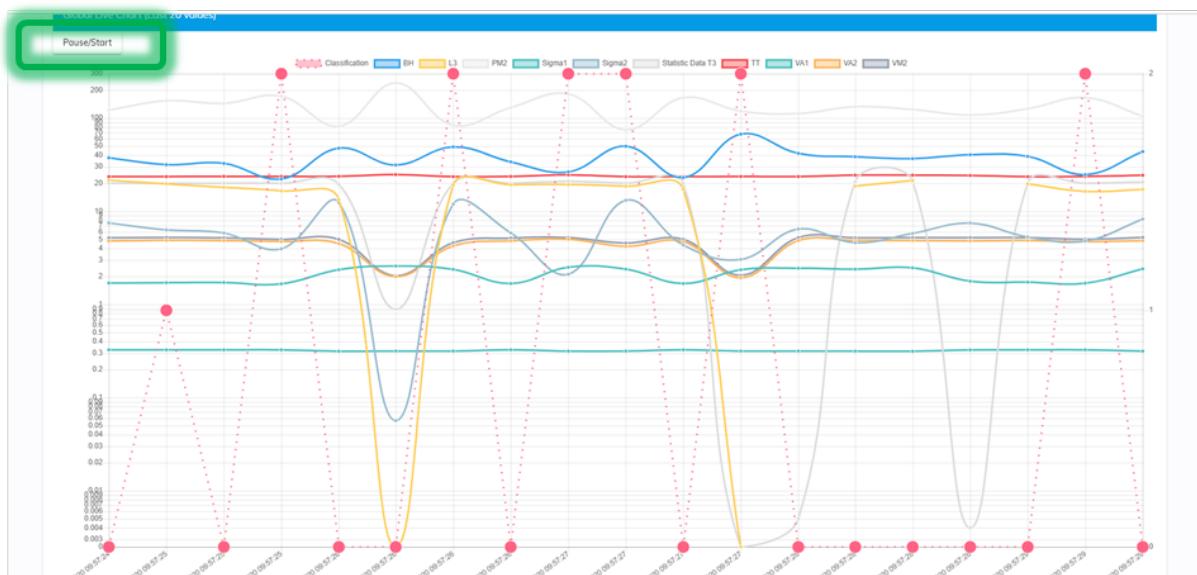
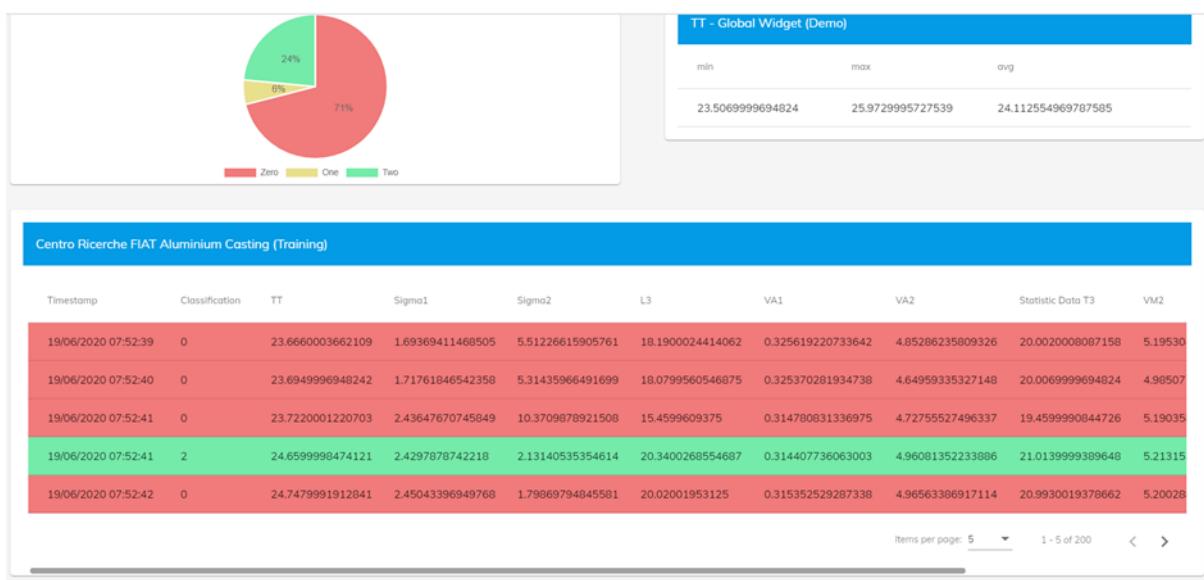
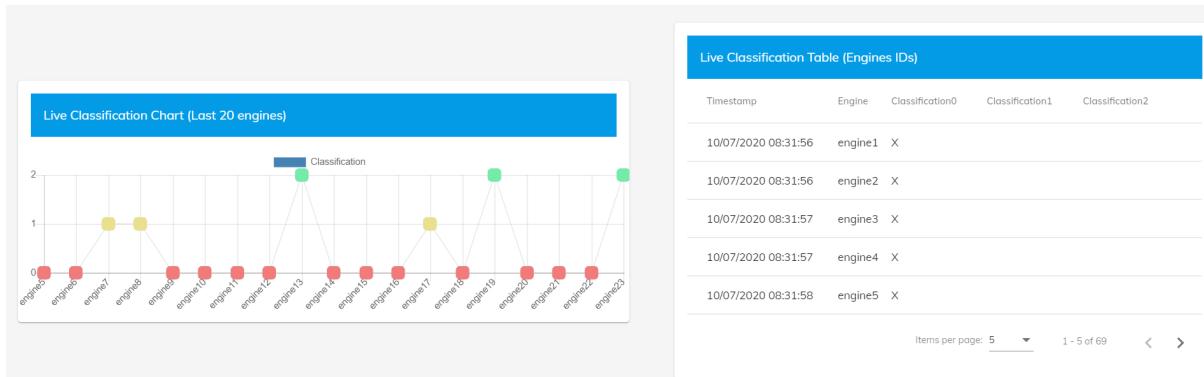


Figure 24: Parameters trend

A data table shows a colour-coded presentation of the results according to a given dimension of data (classification level).

**Figure 25: Classification level**

Furthermore, it is possible to quickly visualise in real time the sequence of engine blocks with their level of classification.

**Figure 26: Classification levels versus engine produced**

In this case, batch analytics were used to develop the high-level algorithms in order to identify and select the critical parameters, by providing a ‘Co-Develop-Mode’ to timely check the status of the process and classify the quality levels that we have identified as KPIs. Subsequently, CRF connected its internal server to the I-BiDaaS platform through a Virtual Machine created by the I-BiDaaS technologists for sharing data in real time. Every two minutes, corresponding to the production time of an engine block, data are copied in a folder in the Virtual machine, analysed near real time and provide a response in a few seconds. In this way, we can develop a methodology to reduce scrap and waste and prevent repairs and reprocessing, by avoiding unnecessary actions after the die-casting of the engine block, such as impregnation, cooling, storing and management of failed engines.

The possibility to use a Virtual Machine for sharing and copying data was a great solution, provided by the I-BiDaaS consortium that can be easily utilised from industries for which corporate constraints do not allow to share a high volume of data with internal systems and cannot give access to their internal servers.

The following table summarises some of the key points of the use case experimentation:

Table 18: Production process of aluminium die-casting

Experiment #8	Dataset(s): <i>Anonymized Aluminium die-casting</i> <i>Synthetic Aluminium die-casting</i> <i>Thermal Data</i>	Preparation status: <i>Generated 100%</i> <i>Generated 100%</i> <i>Generated 100%</i>	Analytics type: <i>batch/streaming</i>																		
Experiment's Goals	To test the efficiency of I-BiDaaS solution in the context of correlating defects with the production process parameters.																				
Experiment's Questions	<p><i>Q1. What is the quality of the analytics results?</i></p> <p>Q1.1 What is the accuracy of new models with respect to internal CRF Aluminium Casting models?</p> <p><i>Q2. How efficient is the process of data analytics?</i></p> <p>Q2.1 How efficient is the performance of the analytics application (algorithm)?</p> <p>Q2.2 How efficient is the visualisation of the analytics solution to allow a quick intervention with specific actions?</p>																				
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Data selection 2 – Synthetic data preparation <ul style="list-style-type: none"> - generate rules - fabricate synthetic data - upload data set 3 – Real data preparation 4 – Data analysis <ul style="list-style-type: none"> - select algorithm - custom algorithm 5 – Data visualisation 6 – Adjust data fabrication rules																				
Experimental Subjects (participating on any of the steps above)	<table border="1"> <thead> <tr> <th>Role</th> <th>Steps involved</th> <th>No of participants</th> </tr> </thead> <tbody> <tr> <td>Quality assurance and control managers</td> <td>1, 4</td> <td>1</td> </tr> <tr> <td>Data analysts</td> <td>1-4</td> <td>2</td> </tr> <tr> <td>Financial administrators</td> <td>4</td> <td>1</td> </tr> <tr> <td>Infrastructure engineers</td> <td>1</td> <td>1</td> </tr> <tr> <td>IT security personnel</td> <td>1</td> <td>1</td> </tr> </tbody> </table>			Role	Steps involved	No of participants	Quality assurance and control managers	1, 4	1	Data analysts	1-4	2	Financial administrators	4	1	Infrastructure engineers	1	1	IT security personnel	1	1
Role	Steps involved	No of participants																			
Quality assurance and control managers	1, 4	1																			
Data analysts	1-4	2																			
Financial administrators	4	1																			
Infrastructure engineers	1	1																			
IT security personnel	1	1																			

3.3.5 Generic experiments

Generic experiments aim to evaluate the functionality and usability of the I-BiDaaS platform from the perspective of potential generic end-users. These experiments are meant for the platform usage beyond the concrete industrial use cases defined within the project, aiming at a wider solution usability and applicability. To this end, Experiment #9 shown in Table 19 corresponds to the ‘Self-Service mode’ of the platform and is targeted to non-experts and experiment #10, shown in Table 20, is for the expert users (PyCOMPSS developers).

End-to-end solutions have been defined to offer comprehensive systems in alignment with I-BiDaaS infrastructure solutions, considering that an end-to-end solution may cover everything from the setup of the project, the selection of data sources, the selection, setup and execution of the algorithms until the visualisation of results.

Table 19: Experiment definition for end-to-end I-BiDaas platform in self-service mode

Experiment #9	Dataset(s): <i>Synthetic datasets used for proof of concept</i>	Preparation status: <i>Generated 100%</i>	Analytics type: <i>Batch</i>
Experiment's Goals	To evaluate user experience of the platform in self-service mode. Experimental subject: Data analysts		
Experiment's Questions	<p><i>Q1. How easy and intuitive did you find using the I-BiDaas platform when it comes to (project setup, source selection, algorithm selection and setup, results visualisation)?</i></p> <p><i>Q2. How satisfied are you with your I-BiDaas experience today (user guidance and usability, Information output, overall impression)?</i></p> <p><i>Q3. Did you experience any crash or malfunction while using the I-BiDaas platform?</i></p> <p><i>Q4. How would you rate the potential of I-BiDaas solution in the following aspects (operability, innovation, compliance, privacy awareness, cost reduction)?</i></p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Project setup 2 – Selecting a data source 3 – Algorithm selection and setup (run the user questionnaire, use the interactive guidance to select an algorithm, set the amount of resources to be used) 4 – Execute the algorithm 5 – Results visualization		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Data analysts	1-5	10-20

Table 20: Experiment definition for end-to-end I-BiDaas platform in expert mode

Experiment #10	Dataset(s): <i>Bank transfer tokenized data</i>	Preparation status: <i>Generated 100%</i>	Analytics type: <i>Batch</i>
Experiment's Goals	To test usability of the platform in expert mode. Experimental subject: Expert Big Data PyCOMPSs developer		
Experiment's Questions	<p><i>Q1. How does the use of the platform impact user productivity?</i></p> <p><i>Q2. How do expert Big Data COMPSs developers rate usability of the I-BiDaas platform in the expert mode?</i></p> <p><i>Q3. To what extent does the platform correspond to the user needs?</i></p> <p><i>Q4. Cost/Effort reduction (because of not having to maintain or setup a PyCOMPSs environment for development)?</i></p>		
Experimental Workflow (based on the generic workflow, to be further refined)	1 – Create a new project 2 – Select a data source 3 – Upload code for the desired algorithm, optionally using a template provided 4 – Edit the code and re-upload		

	5 – Execute the algorithm 6 – Inspect resources used 7 – Results retrieval		
Experimental Subjects (participating on any of the steps above)	Role	Steps involved	No of participants
	Expert Big Data PyCOMPSS developer	1-8	2-5

3.3.6 Cross-sectorial experiments

Several discussions have taken place between CAIXA and TID for finding potential cross-sectorial experiments that help in the evaluation of I-BiDaaS. Most of the use cases that were considered require to share very sensitive data from CAIXA or TID customers. CAIXA studied the way to extract data without breaking data privacy and perform certain level of big data analytics. However, the results from those analytics are evaluated internally after decrypting the data.

Other encryption mechanisms that could be used to do analytics mixing data from different entities were evaluated (i.e. FHE) but we arrive at the conclusion that only very simple operations can be done with this kind of encryption mechanisms, so they are not viable either (as described in previous section 3.3.3.4).

Therefore, the only research line that was found in order to correctly perform the cross-sectorial experiments and overcome the data privacy barriers set by the General Data Protection Regulation (GDPR) was going into the Federated Machine Learning (FML) direction. FML allows different entities to perform big data analytics with the sensitive data they own and do it on their premises. Models from their data are generated by each entity and they are combined in a common macro-model constructed from the data models from each entity. This common model should be previously defined and agreed by the entities and it should have some parameters in common defined by the different entities in order to be useful. After constructing the macro-model, it can be consulted by all the entities in order to enhance each entity model and extract new insights not previously available.

We consider FML a very interesting approach to work on cross-sectorial use cases in which we want to mix sensitive data that, by regulation, cannot be shared between industrial entities. We identified this approach as a hot research topic that is trending up. However, further experimentation with this approach was considered out of the scope of the I-BiDaaS experimentation because of its complexity, and it would require newer initiatives and projects that focus on this research line.

However, we need to highlight that during the course of the project, we performed extensive analysis on how the I-BiDaaS solution can support cross sectorial experiments. I-BiDaaS offers both experts and non-IT experts, all the necessary technologies to take, to understand, to process, to visualize data and to extract value from data as long as the data become available.

4 Experimental Evaluation

4.1 Overview

This section describes the evaluation process of the I-BiDaaS solution, covering the analysis of the different tools and modules provided and the overall performance of the I-BiDaaS prototype. It follows the evaluation methodology followed in D1.3 [3], defining and reporting smooth and adequate running of the experiments according to the experimental protocol and demonstrating how I-BiDaaS solution can effectively aggregate, pre-process, manage and synthesize different types of data, noisy and large-scale data sets in both batch and real-time processing.

The evaluation process provides structured feedback to the development process both from the data providers and the technology owners in order to ensure the project's impact, thus fostering platform's long-term sustainability.

4.2 Data Quality Evaluation

4.2.1 Data quality from the perspective of assessing algorithm scalability

In this section, we provide a report regarding testing the data quality with respect to scalability or to be more specific, with respect to algorithm scalability in testing with synthetic data. This means that, in order to measure synthetic data quality, we need to see how it behaves (scaling-wise) compared to real data. Ideally, the differences in scaling should be minimal.

We tested this hypothesis on the CAIXA's dataset that corresponds to the use case '*Analysis of relationships through IP addresses*'. The real (tokenized) dataset has 295838 samples, while the dataset generated with IBM's TDF tool has 481672 samples. We took 5000 samples from both of these datasets and tested a K-Means implementation from the scikit-learn library¹⁸ on them.

Our conclusion is that the synthetic data behaves very similarly to the real data regarding the execution time/CPU scaling, which proves that it is suitable for algorithm scalability testing.

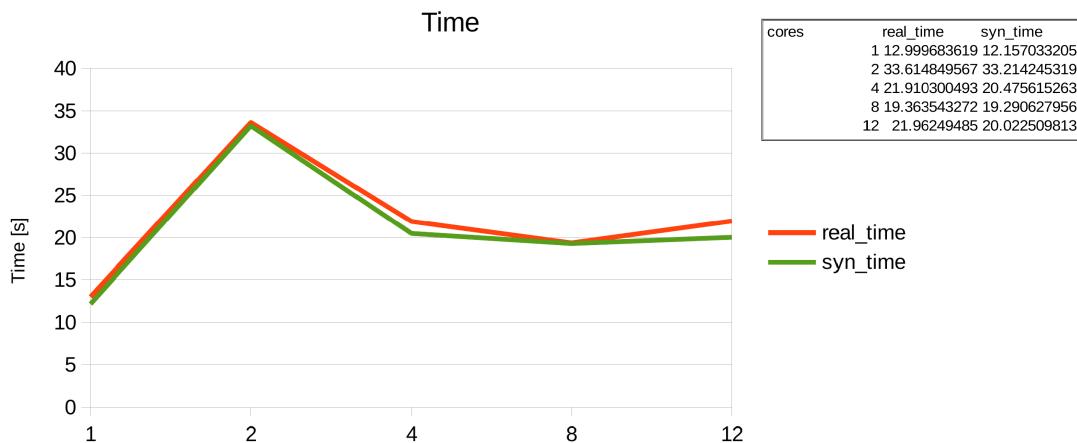


Figure 27: Execution time versus number of cores for the real (tokenized) dataset (red line) and the TDF-generated synthetic dataset (green line) for the CAIXA's use case ‘Analysis of relationships through IP addresses’.

¹⁸ <https://github.com/scikit-learn/scikit-learn>

4.2.2 Specific and general utility

This evaluation was carried out, in several phases, over data that were fabricated for the CAIXA use case '*Analysis of relationships through IP address*' and the CRF use case '*Production process of aluminium die-casting*'. Both use cases include detailed data definitions, specifications and descriptions which can be found in D2.1 [4]. The synthetic data was fabricated by TDF using these definitions.

The TDF synthesizer accepts data description rules (constraints defined by the user) and fabricates data that satisfies all the constraints (using a solver). Although the synthesized data is guaranteed to satisfy each of the provided constraints, there is no assurance that the TDF synthesizer model corresponds to the model that generated the original data. When the generative model adheres too closely to the proposed utility model, the validity checks such as the existence of other interactions may not be apparent in the synthesized data. That is why general and specific measures of utility are required in providing an assessment for the synthetic data model. For that end, initially, the data provides performed a set of tests to validate the structure as well as a set of minimal requirements of internal applications that make use of such data.

To extend analyses and experiments, IBM performed a generic evaluation process for the real data (provided by the use case providers) compared with the fabricated data. This evaluation is concerned with methods to judge whether the fabricated data have a distribution that is comparable to that of the original data, what is commonly referred to in the literature as general utility. In addition to the general utility, we also consider specific utility, i.e. the similarity of results of analyses from the synthetic data and the original data.

As a general measure of data utility, we used the propensity score mean-squared-error (pMSE), to the specific case of synthetic data. As specific utility measures, we used confidence intervals overlap and standardized difference in summary statistics, which we added to the general utility results.

Specific Utility

Synthetic data utility is often assessed by analysis-specific measures which compare data summaries and/or the coefficients of models fitted to synthetic data with those from the original data. If inferences from original and synthetic data agree, the synthetic data are said to have high utility. Published evaluations of synthetic data using specific utility measures, usually for just a few selected analyses, have highlighted differences in the quality of syntheses.

We applied data analysis over the real and synthetic datasets that included the inference of single and multi-attribute constraints and compared a selected subset of the results. The single attributes inferred constraints include best fitted distributions (and the corresponding parameters), min-max values, value frequencies, patterns, formats and other statistical properties. The multi-attribute inferred constraints included value correlations between tuples (up to size 3) of columns (Numeric, categorical, dates and polynomial relations of up to degree 4), as shown in Figure 28.

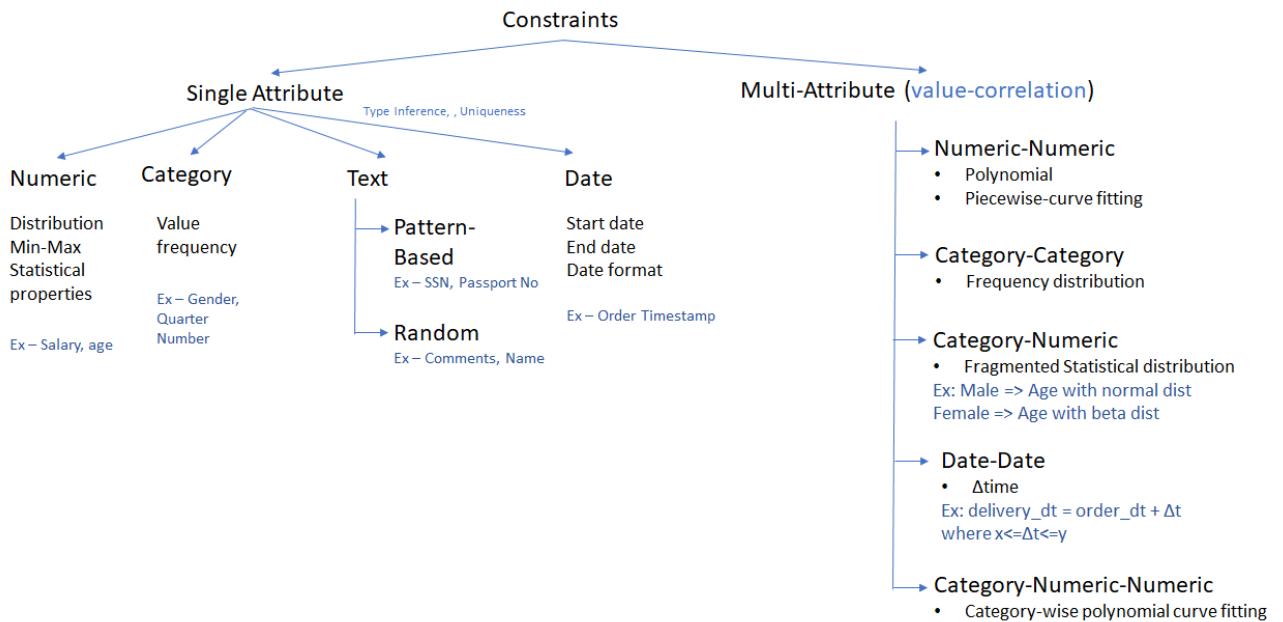
**Figure 28: Data attributes analysed for single and multi-columns**

Figure 29 shows an excerpt from the data analysis results of a single column (PM2 parameter) in the ‘Production Process of Aluminium die-casting’ dataset, in which there are 187 columns that contain the main parameters identified for the detection of the quality level KPI.

```
{
    "col_index": 7,
    "col_name": "PM2",
    "datatype": "float",
    "num_missing": 0,
    "is_unique": false,
    "num_unique_values": 369788,
    "min": 140.0001220703125,
    "max": 247.9962615966797,
    "mean": 184.0800018310547,
    "standard_deviation": 32.16999816894531,
    "min_diff": -107.52392578125,
    "max_diff": 107.23143005371094,
    "mean_diff": -0.00015100765359116124,
    "monotonic": false,
    "distribution": [
        {
            "distribution_name": "norm",
            "parameters": {
                "loc": 184.0800018310547,
                "scale": 32.16999816894531
            },
            "p_value": 0,
            "min_value": 140.0001220703125,
            "max_value": 247.9962615966797,
            "num_rows": 499999
        }
    ],
    "is_category": false
},
```

Figure 29: An excerpt of a single column (PM2)

General Utility

Previous work has suggested various general measures of utility for data that have undergone disclosure control. Generally, these measures consider the distributional similarity between the original and fabricated datasets, with greater utility attributed to masked data that are more similar to the original data. In the broadest sense, measures such as distance between empirical Cumulative Distribution Functions (CDFs) or the Kullback-Liebler (KL) divergence give an estimate of difference. Karr et al. (2006) [16] and the follow-up paper Woo et al. (2009) [17] discussed and implemented various distributional measures such as the KL divergence, an

empirical CDF measure, a method based on clustering, and one that uses propensity scores to estimate general utility. They compared these measures for micro-aggregation, additive noise, swapping, and resampling methods, and they evaluated the propensity score method as the most promising. Propensity scores represent probabilities of group memberships, commonly used in causal inference studies. To use them as a measure of utility, there is a need to model group membership between the original and the masked data to get an estimate of distinguishability where small distinguishability relates to high distributional similarity between the original and masked data. If the propensity scores are well modelled, this general measure should capture relationships among the data that methods such as the empirical CDF may miss.

The propensity score method, given in Woo et al. (2009) [17] can be summarized as follows. The ‘n’ rows of the original and ‘m’ rows of the synthetic data are merged with the addition of a variable ‘I’ which indicates the source of the data (0 – real and 1- synthetic). A propensity score P_s is estimated for each of the $n + m$ rows, as the probability of classification for the indicator variable, using predictors based on the variables in the data. The difference (MSE) between these estimated probabilities and the true proportion of records ($\frac{m}{n+m}$) from the synthetic data in the merged data gives the utility statistic

$$pMSE = \frac{1}{(n+m)} \sum (P_s - \frac{m}{n+m})^2$$

The method can be thought of as a classification problem where the desired result is poor classification (50% error rate), giving better utility for low values of the pMSE.

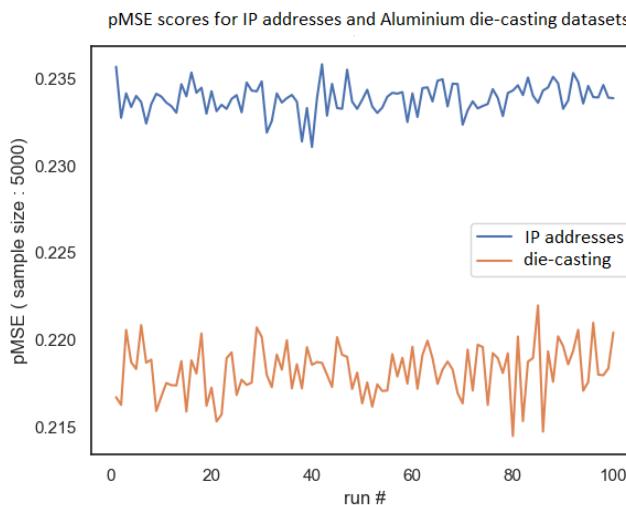


Figure 30: Results from 100 random sampling taken from the real and the synthetic data (5000 datapoints of each) with the pMSE calculated from a logistic model

Randomly sampling 5000 datapoints from the real and synthetic datasets, and using a logistic regression to provide the probability for the label classification we were able to show that the measured mean pMSE score for the CAIXA IP dataset is 0.234 with a standard deviation of 0.000835, and the measured pMSE score for the ‘Production Process of Aluminium die-casting’ dataset is 0.218 with a standard deviation of 0.00146.

4.3 The I-BiDaaS integrated solution and architecture implementation

I-BiDaaS platform takes into consideration many important features that have been implemented in order to utilise the I-BiDaaS solutions developed by data analysts and data technologists. Batch and streaming analytics have been enhanced respectively through the implementation of more data, high-level algorithms, better testing and through analysing real-time data via complex event processing.

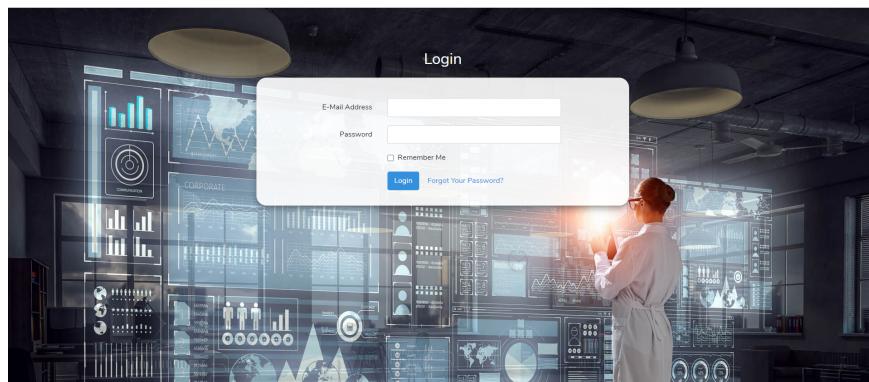


Figure 31: Login to I-BiDaaS platform

After the login, the end-users visualise a dashboard in which they can select how to use the I-BiDaaS solution with 3 possible modes of operation, with different profiles depending on their knowledge or business needs, as shown in the following figure.

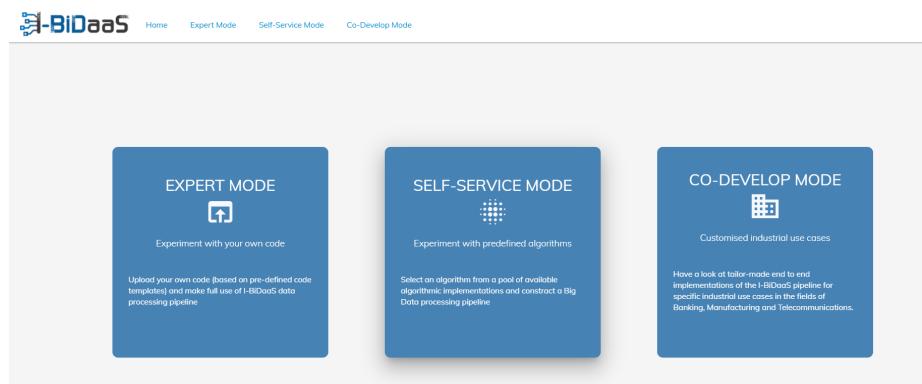


Figure 32: Different modes supported by the I-BiDaaS platform

- The Expert mode allows experts (developers) to upload their own data analytics code based on the available I-BiDaaS highly reusable templates.

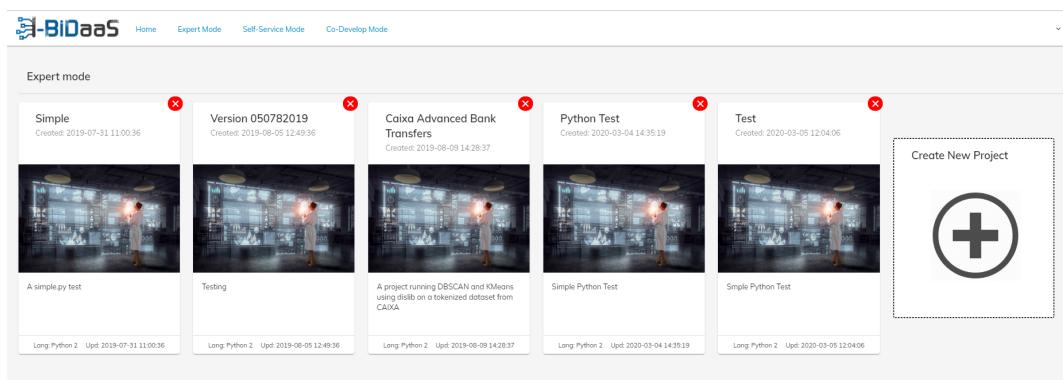


Figure 33: Expert mode

- b) The Self-service mode allows users that have the relevant domain knowledge and some knowledge about data analysis (non-experts) to easily construct Big Data pipelines in a user-friendly way, selecting a pre-defined data analytics algorithm from an available list.

The screenshot shows the I-BiDaaS interface in 'Self-service mode'. It displays four cards representing different machine learning algorithms:

- LASSO ADMM**: Least Absolute Shrinkage and Selection Operator algorithm for regression analysis, solved in a distributed manner. The LASSO model uses L1 regularization to induce sparsity and prevent overfitting of the model.
- K-Means - Prediction**: The objective of K-means is to group similar data points together in a user-specified number of clusters (K).
- Random Forest**: Random forests is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees of training data and outputting the class (classification) or mean prediction (regression) of the individual trees.
- Support Vector Machine**: In machine learning, support-vector machines (SVMs), also support-vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Each card includes a status indicator (Status: open) and an update timestamp (Upd: 2019-07-19 08:38:24, Upd: 2019-07-19 08:39:46, Upd: 2019-07-19 08:39:46, Upd: 2019-07-19 08:39:46). A modal dialog box at the bottom asks 'Which algorithm should you select?' with a 'Start' button and a text input field asking 'Please answer the following questions to get a recommendation on which algorithm to use based on your data'.

Figure 34: Self- service mode

- c) The Co-Develop mode corresponds to an end-to-end solution for a given industry project developed by the I-BiDaaS team (the I-BiDaaS use cases). Continuing from M19, when the 1st integrated solution was released, the I-BiDaaS integrated solution was expanded to include all the use cases provided by CAIXA, CRF and TID.

The screenshot shows the I-BiDaaS interface in 'Co-Develop mode' displaying seven use case cards:

- Caixa IP Address Relations - Batch processing**: The CAIXA IP Addresses Use Case is about extracting relations between IP addresses. In this use case batch processing of big data emerges as an innovative tool that can be instrumental in fraud detection.
- Caixa IP Address Relations - Stream processing**: The CAIXA IP Addresses Use Case in streaming mode allows for real-time IP relationship detection on incoming data and also the graph representation of relationships among the IPs.
- Centro Ricerche FIAT Aluminium Casting (Training)**: The use case deals with monitoring the quality of the production process of aluminium casting. The goal is to use Big Data to monitor the quality of the production process and operational efficiency; in particular, the quality issues on the automotive component.
- Telefonica CC Sentiment - Stream Processing**: The use case is about monitoring the sentiment of calls to Telefonica's call centres. Using GPU accelerated processing, it monitors the evolution of sentiment over frequent words in calls, better analysis and increased performance are envisaged.
- Telefonica Mobility Data - Locations Map**: This case concerns the monitoring of the load percentages of Telefonica's antennas over a period of time in order to observe and therefore to predict possible congestions caused by the users' movement around the area.
- Telefonica Antenna KPIs - Stream Processing**: This use case demonstrated high accuracy predictions of Telefonica's antennas that will become the next 'hotspots', i.e. will have increased load. Monitoring these anomalies can help operators perform predictive maintenance tasks and avoid senior breakdowns.
- Centro Ricerche FIAT Predictive Maintenance (Batch Processing)**: In this use case, sensor data have been analysed in order to identify anomalies in the measured values. Monitoring these anomalies can help operators perform predictive maintenance tasks and avoid senior breakdowns.

Each card includes a status indicator (Status: open) and an update timestamp (Upd: 2018-12-12 10:33:08, Upd: 2018-12-12 12:23:52, Upd: 2018-02-26 09:23:03, Upd: 2018-02-26 09:23:03, Upd: 2018-02-26 09:23:03, Upd: 2018-02-26 09:23:03, Upd: 2018-02-26 09:23:03).

Figure 35: Co-Develop mode

The I-BiDaaS architecture is depicted in the following figure.

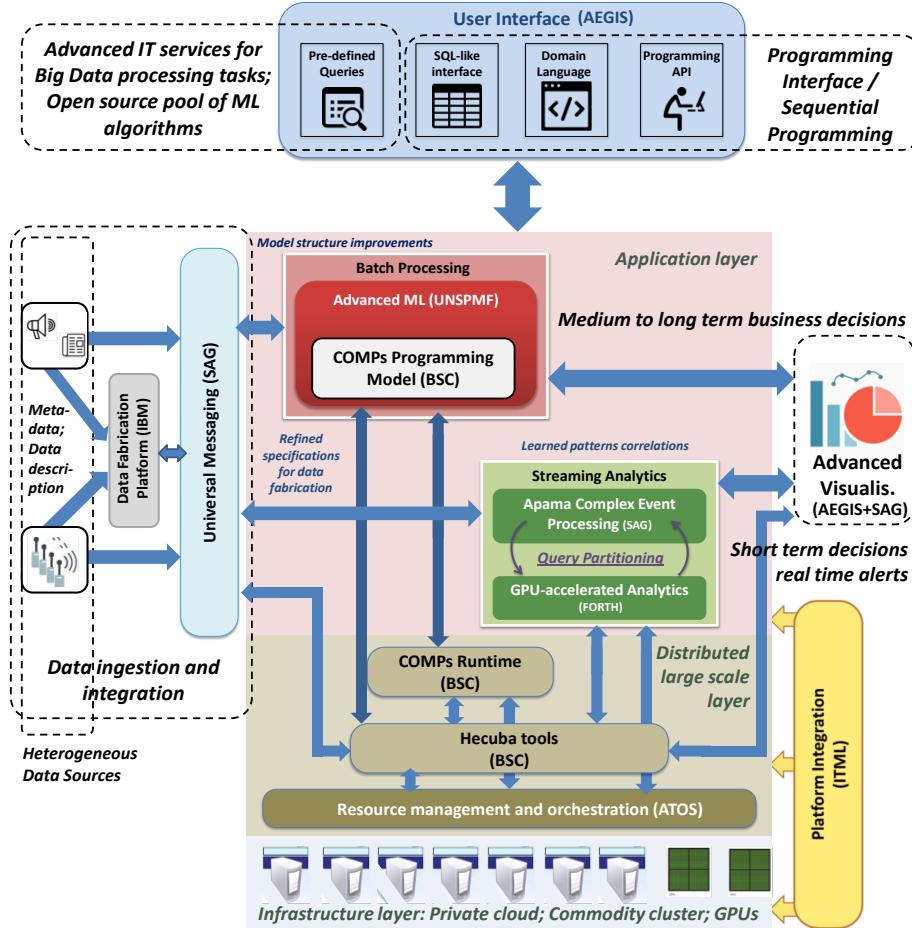


Figure 36: The I-BiDaaS architecture

The system architecture is shown in Figure 37. The cluster that runs the batch processing jobs is based on docker swarm¹⁹. The swarm consists of a manager node and a set of worker nodes. The orchestrator assigns dynamically a set of workers from the set of available nodes in docker swarm, based on the user's preferences and set-up inserted through the UI. These workers exploit the Cassandra DB and the shared FS in order to complete the requested job.

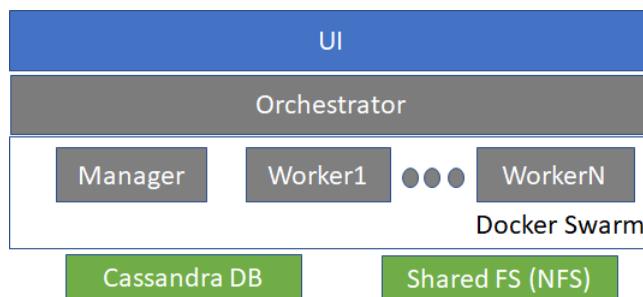


Figure 37: System architecture: The orchestrator and the docker swarm

¹⁹ <https://docs.docker.com/engine/swarm/>

For the case of streaming analytics, the architecture is presented in Figure 38. In this case, the orchestrator collaborates with the Universal Messaging bus provided by SAG that collects the data and feeds the APAMA analytics engine.

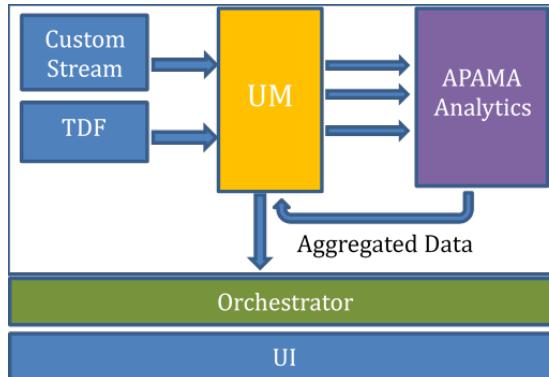


Figure 38: System architecture: The orchestrator and APAMA analytics

4.4 Experiments verification and validation

Experiments verification and validation corresponds to the *Operation* step of the I-BiDaaS experimental process (described in deliverable D6.1 [19]), aiming at the evaluation of the I-BiDaaS platform according to the experiments' definition and against the stakeholders' requirements. A main characteristic of the I-BiDaaS experimental process is that it considers both technical and business requirements. In particular, it aims to evaluate both the performance of the I-BiDaaS solution, but also its alignment with the needs of the industrial users. To this end, the experiments verification and validation integrates the following aspects:

- Quantitative (technology-centred) evaluation: Evaluation of the quality of the I-BiDaaS platform in parts and as a whole using appropriate benchmarks. The main stakeholders in this phase are technology providers.
- Qualitative (user-centred) evaluation: Experimental evaluation of the I-BiDaaS platform in a real business setting and against user requirements. The main stakeholders in this phase are business users (data analysts, financial administrators, etc.), as well as other users such as IT administrators, Big Data developers, etc.

To this end, a set of indicators has been defined against which validation of the I-BiDaaS solution efficiency can be measured. Such indicators reflect both technology features at component and platform level (e.g., operational performance), as well as business key performance indicators (e.g., service quality, time efficiency). The former are use case independent, whilst the latter reflect the specific needs expressed in each use case. Alignment of both types of indicators has been a main objective of the experimental definition phase.

For each indicator, a set of quantifiable metrics has also been defined, whose measurement relates to the achievement (or not), of a specific indicator. For example, in the case of operational performance, relevant metrics include execution time and throughput. In the case of service quality, a relevant metric might be the 'Overall Equipment Effectiveness (OEE)'. For the metrics related to technology indicators, available big data benchmarks can also be used. An initial investigation of applicable big data benchmarks has been performed during the I-BiDaaS baseline phase and reported in D1.3 [3]. This has been further informed from recent research in the area, reported in the results of ongoing projects, such as the classification of big

data benchmarks proposed by the European DataBench²⁰ project, in which some of the most well-known benchmarking tools are classified according to the benchmark categories, type, domain, data type and metrics measured.

4.5 Quantitative evaluation

The quantitative (technical) evaluation focuses on the evaluation of quality of the I-BiDaaS platform in parts and as a whole through testing using appropriate benchmarks, where available. This section continues the work reported in the D6.3 [2] concerning the evaluation of the results of I-BiDaaS, by providing an update of the individual parts and the overall I-BiDaaS solution evaluation at M32.

4.5.1 Individual parts evaluation

The progress regarding the quantitative evaluation of the I-BiDaaS solution of each platform component is reported in this section in order to provide an update of the evaluation indicators used for the verification and validation of the I-BiDaaS solution in parts. In Table 21, measurement obtained for each of the module, developed in I-BiDaaS, are reported:

Table 21: Component evaluation

Indicator	Metric	Benchmark (if applicable) ²¹	Measurement obtained
TDF (IBM)			
Scalability	Linear (in number of TDF instances) speedup in generated records	This version runs a parallel CSP solver	Data generation was performed on a multi core Virtual Machine (8 CPUs, 4GB RAM) and speedup was linear
Validity	Generated data must fit the data model	The data fits the model	Data fits the model
Performance	Number of generated records per time unit	Average of 52 records per sec	Average of 52 records per sec was measured on the data project of the Production process of Aluminium die-casting
Accuracy	Measured against real data	Benchmark the quality of the fabricated data by applying analytics on both real and fabricated data.	Data was manually examined by relevant partners, compared by applying data analysis on both datasets (specific utility) and a 0.032673244 pMSE score was calculated for the CAIXA IP transactions dataset
Availability	No of crashes	No crashes	No crashes observed during the runs performed
Universal Messaging (SAG)			
Scalability	Response time	seconds	ms
	Data throughput	MB/second	100 – 500 MB/s depending on message length
	Resource utilisation	MB and CPU %	2-4 GB memory 4-5% CPU

²⁰ <https://www.databench.eu/>

²¹ For the non-applicable: Units are included instead of benchmarks.

Operational performance	Execution time	seconds	ms
	Latency	seconds	ms
Reliability	Data failure	N/A	no failure
Compliance	Measured against relevant standards	JMS, MQTT, AMQP	in line with standards
Advanced ML (UNSPMF)			
Quality of results	For optimization solver: Objective function value; constraints violation value	Standard solver on the same problem, if feasible for the given data set and if implementation is available: CVXPY ²² , relevant MPI implementation. Targeted to have a small deviation with respect to the benchmark, e.g., $10^{(-2)}$	See Table 6 and the associated description in D6.3 [2]
	For supervised learning: Training error; testing error; classification accuracy	Benchmark: Sklearn ²³ , for moderate size data. Target: have a comparable result with the benchmark, with improved scalability.	See Tables 5-7 in D3.1 [20]; Table 7 in D6.3 [2]; and Subsection 4.1.1.3 and Figure 13 in D3.3 [6], and the associated descriptions.
	For clustering: silhouette score	Benchmark: Sklearn, for moderate size data. Target: have a comparable result with the benchmark, with improved scalability.	See Tables 9 and 10 in D6.3 [2] and the associated descriptions.
Scalability	Execution time versus number of nodes (cores)	Respective MPI implementation, if available. Target: have scalability which is comparable with MPI (lower performance than MPI expected due to using COMPSs framework with less programming and system optimization effort.)	See Tables 11-14 in D6.3 [2], and Figure 13 in D3.3 [6], and the associated descriptions.
Performance	For testing a novel algorithm: Number of iterations; number of messages exchanged between the nodes	Comparison of the communication-efficient distributed stochastic gradient method proposed therein and compared it with several state-of-the-art methods in an emulated environment with respect to the number of iterations and number of exchanged messages. The proposed algorithms show comparable performance iteration-wise while significantly reducing the number of exchanged messages.	See reference [6] and Subsection 4.1.1.3, second paragraph of D3.3 [6]
CEP Engine (SAG)			
Scalability	Response time	seconds	ms

²² Python-embedded modelling language for Convex optimization problems, available at https://www_cvxpy.org/

²³ <https://scikit-learn.org/stable/index.html>

	Data throughput	MB/seconds	up to 500 MB/s
	Resources Utilisation	MB and CPU %	1-2 GB memory 3-4% CPU
GPU-accelerated pattern matching (FORTH)			
Performance	Throughput	Filtering data using pattern matching in TerracottaDB using GPU-acceleration. Baseline: Filtering data using pattern matching in vanilla TerracottaDB.	The throughput ranges between 1.25 - 1.43 Mtuples/sec when processing tuples of name/surname pairs, using a GTX 980 graphics card (more details on D6.3 [2]).
	Latency		The latency ranges between 0.07 - 0.08 sec when processing tuples of name/surname pairs, using a GTX 980 graphics card (more details on D6.3 [2]).
Resource management and orchestration module – RMO (ATOS)			
Operational performance	Response time	- Page response time (in seconds) - Transactions processed - Network error, latency & utilization	100-150 ms
	Resource utilization	Delay between a client request and a cloud service provider's response. (in seconds)	< 1 second
Scalability	Average of assigned resources among the requested resources	Makespan of the service creation until the deployment of the resources are acknowledged. (in seconds)	The range of the deployments tested goes from 33 seconds for the deployment of a single VM using a small flavor (vcpu:1, disk(GB): 20, mem(MB): 2048) to 150 seconds for more complex deployments based on Kubernetes, which includes large flavours (vcpu:4, disk(GB): 80, mem(MB): 8192) for a master, worker and balancer nodes.
Availability	Responsiveness	Verify that the number of resources is between the resource limits (max & min) as defined in the blueprint.	TOSCA specs allow us to define the min and max number of instances allowed for each resource type that compose our cloud-based service, as well as the relationships between them.
Reliability	Service Constancy	How much time the service provider guarantees that your data and services are available (in percentage)	around 90%
	Accuracy of Service	Rules defined to ensure service reliability (number of replicas and policy types)	The scale-up and scale-down boundaries can be defined, by including the minimum requirements for a service to operate as well as how much each of the resource types can scale
	Fault Tolerance	Ability to continue providing service after a failure	The resources that compose the service are continuously

			monitored in order to react in case of failure
Visualisation Tool (AEGIS)			
Operational performance	Response time	7-10s	Initial Page load: 9.92s (avg) Internal Page Loading: <1.25s (avg)
Availability	Uptime	95%	No downtime was experienced during the operation of the platform. Platform was only not available during planned maintenance and version updating.
Reliability	Fault Tolerance	Interface responsive in case of data errors. Informative messages to users.	No errors were detected that prevented interface from being responsive. Faulty information is not permitted by the interface.
Usability (Efficiency)	Perception of task completion quality	>80% positive perception. Results via 1-5 scaled questions.	87.5% achieved
Usability (Satisfaction)	Degree to which user needs are satisfied - look and feel	>80% positive perception. Results via 1-5 scaled questions.	87.5% achieved
Qbeast (BSC)			
Effect on Machine learning algorithms	Speedup	Benchmark: Dislib and MLlib ²⁴ Baseline: Dislib and MLlib without using Qbeast	Qbeast improves after multiple Read-Optimizations for three different types of queries (All 0.01%, Olfactory 1% and Inhaler 1%). Different speedup can be achieved in the three queries, ranging from 24.51 X improvement to a “mere” factor 2.37. In query “All 0.01%”, we have the highest speedup as we benefit the most from the efficient sampling of Qbeast (see Table 22)
Scalability		Benchmark: synthetic geographical queries Baseline: PostGis ²⁵	For I/O, both Cassandra and Qbeast perform very similarly, and improve 80% when doubling the nodes. The scalability is not linear, as the replica is synchronous, which adds latency and increases resource usage (see Figure 39)
Operational performance	Response time		Qbeast always outperforms PostgreSQL and GPFS in our tests for Net I/O time. For pure response time, Qbeast beats PostgreSQL between 6.5 and 260 factor, depending on the case (see Figure 40-41)

²⁴ The Apache Spark's scalable machine learning library, available at <https://spark.apache.org/mllib/>

²⁵ A spatial database extender for PostgreSQL object-relational database, available at <https://postgis.net/>

	IOPS		With two nodes, Qbeast, achieves 83K IOPS (performing as Cassandra), and approximately improves 80% when doubling the nodes (see Figure 39)
	Disk usage		In terms of disk usage, older versions of Qbeast required to replicate each item 5.29 times on average, while the newest version requires only 1.14.
Availability	% timeouts		0%
Reliability	Fault tolerance on down nodes (amount of missed data and slowdown during the recovery process)		Fault tolerance mechanisms provided by Cassandra (e.g. replication)
Hecuba DBS (BSC)			
Scalability	Speedup	Benchmark: Dislib ²⁶ Baseline: Dislib using files	2.54 max (for K-means)
Operational Performance	Response time		68 % from the original (in KNN)
	IOPS		N/A
Availability	Disk usage		N/A
	% timeouts		0%
Reliability	Fault tolerance on down nodes (amount of missed data and slowdown during the recovery process)		Fault tolerance mechanisms provided by Cassandra (e.g. replication)

Qbeast

During the project, the tests initially proposed for Qbeast have been reconsidered, as described in the previous table. New tests show better the real performance of Qbeast, not only for synthetic benchmarks, but also for real applications. We ran our tests at the Barcelona Supercomputing Center, in MareNostrum IV supercomputer. Each server contains two sockets with an Intel Xeon Platinum 8160 24C for a total of 48 cores and 96 GB of ram for each server. Nodes are interconnected by a 100 GB Intel Omni-Path and a 10 GB Ethernet²⁷. We use the local SATA 240 GB Intel s3520 SSD scratch disk to store data. The disks are rated for sequential reads and write up to 320, and 300 MB/s, respectively, while for random reads and writes up to 65000 and 16000 IOPS. We used a stress tool shipped with Cassandra to benchmark the system, using twice as many machines for the stress tool than the database and performing random insertions with a Gaussian distribution.

In Figure 39, with two nodes, Cassandra and Qbeast perform very similarly, achieving respectively $\approx 84\text{K}$ and $\approx 83\text{K}$ IOPS. Cassandra and Qbeast approximately improve 80% when doubling the nodes. The scalability is not linear as the replica is synchronous, which adds latency and increases resource usage.

²⁶ The Distributed Computing Library, available at <https://dislib.bsc.es/>

²⁷ <https://www.bsc.es/marenostrum/marenostrum/technical-information>

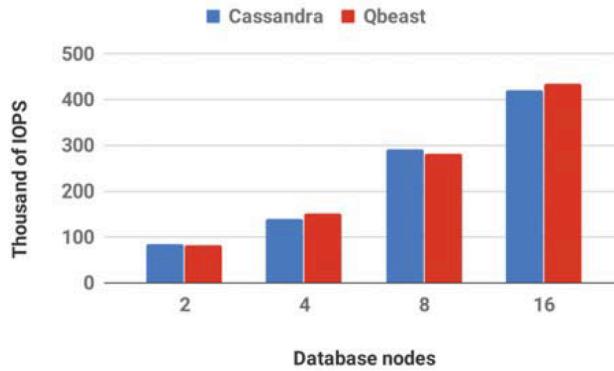


Figure 39: Thousands of IOPS of Cassandra with 2 replicas vs Qbeast

The second part of our experiments uses ALYA²⁸ to test the performance of Qbeast.

Figure 40 reports the net time ALYA spent performing I/O with different backends when increasing the number of workers. We shall note that Qbeast can store and index data faster than the GPFS can write into a not-indexed csv file. Another notable result is that time required for I/O for one Qbeast node or eight is not proportional as the I/O time of ALYA remains approximately constant when varying the number of workers. The Figure also shows that PostgreSQL is considerably slower while ingesting writes and that its speed decreases when increasing the number of concurrent actors.



Figure 40: Net I/O time for 1000 steps with different backends

Figure 41 compares the response time of different configurations of Qbeast and PostgreSQL. We can see how Qbeast always outperforms PostgreSQL.

²⁸ <https://www.bsc.es/research-development/research-areas/engineering-simulations/alya-high-performance-computational>

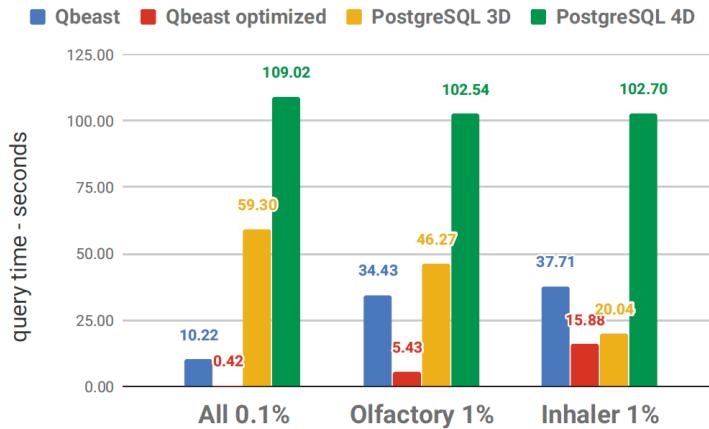
**Figure 41: Response time: Qbeast vs. PostgreSQL**

Table 22 shows how Qbeast improves after multiple Read-Optimizations for three different types of queries (All 0.01%, Olfactory 1% and Inhaler 1%). The table also reports the different speedup we can achieve in the three queries, ranging from 24.51 X improvement to a “mere” factor 2.37. In query “All 0.01%”, we have the highest speedup as we benefit the most from the efficient sampling of Qbeast.

Table 22: Qbeast speedup after few ReadOptimizations

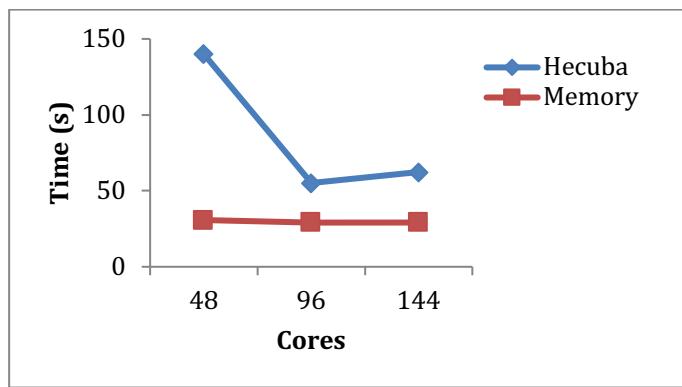
	speedup	RO runs	iterations	cube visited
All 0.01%	24.51	6	2	10
Olfactory 1%	6.34	10	19	19
Inhaler 1%	2.37	8	61	61

Finally, in terms of disk usage, older versions of Qbeast required to replicate each item 5.29 times on average, while the newest version requires only 1.14.

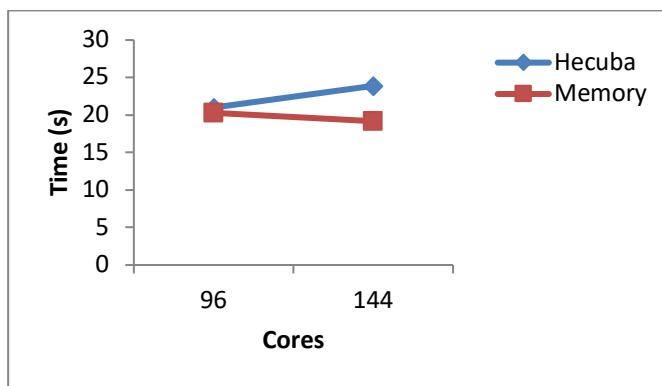
Hecuba

The tests of the integration between Hecuba and Dislib have been executed on MareNostrum IV supercomputer, as well as Qbeast tests. In order to obtain the different measurements, we have performed three different algorithms: K-Means, PCA and KNN.

For the K-Means algorithm, it has been used a dataset of 10 Million samples with 50 features. Dislib gave us the possibility to decide the granularity of the data. To obtain the better performance, we have divided the dataset in 48 blocks which will be operated in parallel, thanks to COMPSs. The performed K-Means is analysing the data to find 50 clusters. We can see that the better performance is obtained when we are using 96 cores, in MareNostrum IV this is equivalent to use 2 computing nodes.

**Figure 42: Response time K-Means**

To perform the PCA algorithm, we have used a dataset of 10 M samples with 50 features, reducing the number of dimensions down to the 3, which is a typically performed PCA in order to do a posterior clusterization and visualization of data. In this case, better performance has been obtained, dividing data in 96 blocks and using 96 cores (2 computing nodes).

**Figure 43: Response time PCA**

As we can see, for the K-Means and PCA algorithm, there is some extra time when using data from Hecuba, this is due to the time necessary to load the block of data, if further updates a cache will be developed, reducing the time difference.

Finally, for the KNN algorithm, a dataset of 10 M samples and 50 features has been used, performing a KNN of 10 neighbors. For computing these tests, the chosen granularity of data has been 48 blocks. As it can be seen the algorithm obtains a better performance when using data from Hecuba, this is due to an improvement on COMPSs functionalities, which detects that data is located on Cassandra, by this way each task can retrieve its own data instead of having to serialize it into a file to pass it to the tasks.

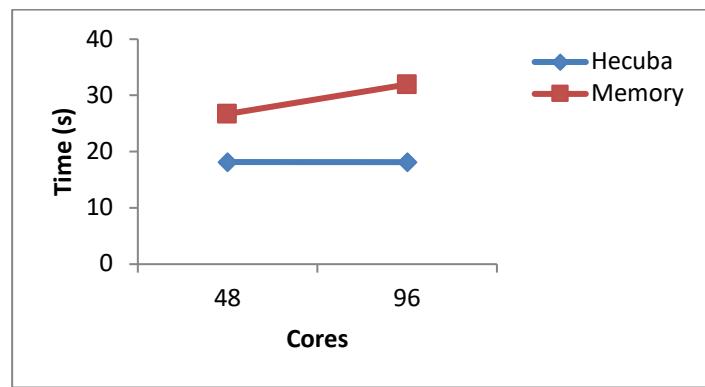


Figure 44: Response time KNN

4.5.2 Overall I-BiDaaS solution evaluation

The following table provides an update of the main metrics used for testing the quality of the overall I-BiDaaS solution, in combination with relevant benchmarks and measurements obtained so far. The measurements are obtained by questionnaires provided in focus groups/webinars organized by Telefonica and CAIXA.

Table 23. Overall I-BiDaaS prototype evaluation results

Indicator	Metric	Benchmark	Measurement obtained
Usability	Task time efficiency	>20% Decrease with respect to current times	5/5
	Perception of time required to accomplish a task	>30% Decrease with respect to current times	4.5/5
	Perception of task completion quality	>80% positive perception. Results via 1-5 scaled questions.	5
Scalability	Speedup		To Be provided in D6.5
Operational Performance	Response time (Latency)		To Be provided in D6.5
	Data throughput (IOPS, no of generated data records per time unit)	Benchmark: Dislib ²⁹ Baseline: Dislib using files	To Be provided in D6.5
	Resources utilization (storage, memory, CPU)		To Be provided in D6.5
Availability	Uptime, % timeout		To Be provided in D6.5
Reliability	Data failure, Fault tolerance		Not measured
Data Security	Compliance with relevant security and privacy regulations and standards	At least 90% compliance	100%
Privacy	Compliance with relevant security and privacy regulations and standards	At least 90% compliance	100%
Compliance	Measured against relevant standards	NA	4.5/5

²⁹ The Distributed Computing Library, available at <https://dislib.bsc.es/>

Cost	Compared against commercial alternatives	>30% reduced costs with respect to competitive commercial solutions	4/5
------	--	---	-----

4.5.3 Tests in relation to I-BiDaaS industry validated benchmarks

Tests in relation to industrial benchmarks will be facilitated through the use of the DataBench Toolbox³⁰, which aims to provide tooling support to Big Data benchmarking users to search, select, deploy and run existing Big Data benchmarks on the one hand, while on the other hand, get the results of the execution, homogenize the technical metrics, and finally help derive business insights and KPIs.

The Toolbox offers ways to reuse existing benchmarks (including HiBench³¹, SparkBench³², BigDataBench³³, YCSB³⁴, TCPx-IoT³⁵, Yahoo Streaming Benchmark³⁶, BigBench/TPCx-BB³⁷) and provides a set of automatisms and recommendations to allow their usage.

To this end, a collaboration between I-BiDaaS and DataBench projects has been initiated in May, aiming to explore how to best exploit the DataBench Toolbox in the context of I-BiDaaS. In Section 5.2, the description of the DataBench webinar, where I-BiDaaS participated, is reported.

4.6 Qualitative evaluation

As presented in D6.3 [2], the definition of the experimental qualitative evaluation follows a goal-oriented approach, whereby for each experiment: first the experiment's goal(s) towards which the measurement will be performed are defined; then a number of questions are formed aiming to characterize the achievement of each goal; and finally, a set of indicators and appropriate metrics is associated with every question in order to answer it in a measurable way. The experimental goals and associated questions for each experiment have been presented in section 3.3. The following sections 4.6.1 - 4.6.4 show the indicators and associated metrics for each experiment. These are defined both at business and application level thus, ensuring (a) that both business and technical requirements are taken into consideration and (b) the traceability among business and application performance.

4.6.1 Telecommunication experiments

Tables 24-26 report the values obtained during the telecommunication experiments with respect to the identified metrics.

³⁰ DataBench Consortium (2019), D3.1 DataBench Architecture, Technical Report, available at <https://www.databench.eu/wp-content/uploads/2020/02/d3.1-databench-architecture-ver.2.0.pdf>

³¹ HiBench Suite, <https://github.com/intel-hadoop/HiBench>

³² SparkBench, <https://github.com/CODAIT/spark-bench>

³³ BigDataBench. <http://prof.ict.ac.cn/BigDataBench>.

³⁴ YCSB, <https://github.com/brianfrankcooper/YCSB>

³⁵ TPCx-IoT, http://www.tpc.org/tpc_documents_current_versions/pdf/tpcx-iot_v1.0.3.pdf

³⁶ YSB, <https://github.com/yahoo/streaming-benchmarks>

³⁷ BigBench, <https://github.com/intel-hadoop/Big-Data-Benchmark-for-Big-Bench>

Table 24: Experiment #1 - Accurate Location Prediction with High Traffic and Visibility

	Indicator	Metric	Baseline value / Benchmark	I-BiDaS value
Business KPIs	Acquisition of insights on the dynamics of cellular sectors	Timely forecasting of mobile phone users movements at scale	A typical large-scale cellular networks can contain any number of cell sites (e.g., 40-100K) that may underperform at any given time; currently no automatic predictive mechanisms are in place	Predict places with high traffic and congestion event
Application Level Performance indicators	Prediction accuracy	Accuracy of forecasting as a function of time, amount of historical data, and prediction horizon	Random prediction	A minimum improvement of 5% (or more) over the random prediction
Platform level performance indicators	Throughput	Minimize the processing time with respect to growing data size, while maintaining real-time delivery of forecasting results for a predefined time window (e.g., 24 hours)	Currently no such solution is in place	Any automated forecasting mechanism that can (reasonably) scale the monitoring of cell sites and their incoming traffic

Table 25: Experiment #2 - Optimization of Placement of Telecommunication Equipment

	Indicator	Metric	Baseline value / Benchmark	I-BiDaS value
Business KPIs	Acquisition of insights on the dynamics of cellular sectors	Timely forecasting cell sites with high traffic (i.e. 'hot spots')	A typical large-scale cellular networks can contain any number of cell sites (e.g., 40-100K) that may underperform at any given time; currently the cell site performance may be done manually using a set of predefined heuristics	Study the spatio-temporal patterns and provide insights on the dynamics of cellular sectors; develop an automatic solution for predicting places with high traffic

Application Level Performance indicators	Prediction accuracy	Accuracy of forecasting as a function of time, amount of historical data, and prediction horizon	Random prediction	A minimum improvement of 5% (or more) over the random prediction
Platform level performance indicators	Throughput	Minimize the processing time with respect to growing data size, while maintaining real-time delivery of forecasting results for a predefined time window (e.g., 24 hours)	Currently no such solution is in place	Any automated forecasting mechanism that can (reasonably) scale the monitoring of cell sites and their performance

Table 26: Experiment #3 - QoS in Call Centres

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Processing costs	Total cost for analysing customer audio calls through a third party.	11,520 calls per year * cost unit, to identify 2,300 low customer satisfaction audio calls.	Analysis of 7,000 low customer satisfaction audio calls x cost unit per year (improved recall for reduced cost).
Application Level Performance indicators	Throughput	% of low customer satisfaction index (CSI) customer audio calls analysed per time unit.	Approximately 2,300 low customer satisfaction audio calls detected (out of 11,520) by human agents (100% recall), per year.	Increase the number of low customer satisfaction audio calls detected by human agents to 7,000 by pre-processing/filtering the audio calls (70% recall).
Platform level performance indicators	Throughput	Number of audio calls processed per time unit.	Given an average call duration of 8.6', a human agent could annotate approximately 6 x 8.6' calls per hour. Assuming a work schedule of 40 hours per week (160 hours per month), this equals to 11,520 calls per year.	The I-BiDaaS platform (configuration with 1 core) will process 12 * 8.6' calls per hour. This equals to 105,120 calls per year.

4.6.2 Banking experiments

Tables 27-29 report the values obtained during the banking experiments with respect to the identified metrics.

Table 27: Experiment #4 - Enhance control of customers to online banking

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Cost reduction	Infrastructure cost	Internal temporal storage cost	Cloud storage cost
	Data accessibility	Number of people accessing data	Order of magnitude of 10	Order of magnitude of 100
	Time efficiency	Time to access data	1 month	1 day
Application Level Performance indicators	End-to-end execution time	Anomalies detected	minutes	seconds
		Time to get analytics results.	minutes	seconds
		Data charging time	1 week	2-days
	Accuracy and reliability of the analytical process	Time to generate business rules.	Domain specific evaluation / Number of Anomalies extracted with commercial product (Data Robot) (order of 10s)	Number of Anomalies extracted with I-BiDaaS (order of 100s)
		Confusion matrix, TP, TN, etc.	No baseline values. The volume of detected and verified fraudulent loggings is not sufficient. Supervised dataset may be built after a first phase of unsupervised analysis of the dataset.	
Platform level performance indicators	Cost	Price of technologies	Cost of commercial product licenses (e.g. DataRobot)	Depending on the selected type of license.

Table 28: Experiment #5 - Advanced Analysis of bank transfer payment in financial terminal

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Time efficiency	Time to access data	2 weeks - 1 month	2 days for creating the dataset.
	Data accessibility	Number of people accessing data	Order of magnitude of 10	Accessing 1-5 data analysts per use case.
Application Level Performance indicators	End-to-end execution time	Data charging time	minutes	1-2 minutes
		Time to get analytics results	minutes	1-5 minutes
		Time to generate business rules	1 week	1-2 days

	Accuracy and reliability of the analytical process	Anomalies detected Confusion matrix, TP, TN, etc.	Domain specific evaluation / Number of Anomalies extracted with commercial product (Data Robot) No baseline values. The volume of detected and verified fraudulent transfers is not sufficient. Supervised dataset may be built after a first phase of unsupervised analysis of the dataset.	1%
Platform level performance indicators	Cost	Price of technologies.	Cost of commercial product licenses (e.g. DataRobot).	Depending on the selected type of license. Order of 100k€

Table 29: Experiment #6 - Analysis of relationships through IP address

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Time efficiency	Time to test new technologies	6 months - 1 year	1-2 weeks
	End-to-end execution time (from data request to data provision)	Time to access data vs. time to generate data	2 weeks - 1 month	1-2 days
Application Level Performance indicators	Reliability and accuracy of the insights generated (the relationships must be valid).	- Accuracy - Recall - TP rate - TN rate - Confusion matrix)	No baseline values. Acceptable rates are 90% of accuracy.	Accuracy 100% (rules already defined in the synthetic data)

4.6.3 Manufacturing experiments

Table 30 and Table 31 report the values obtained during the manufacturing experiments with respect to the identified metrics.

Table 30: Experiment #7 - Maintenance and monitoring of production assets

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Product / Service quality	OEE JPH	92% 18	94.5 % 18.2
	Cost reduction	Maintenance cost	200 k€ every 3 months	100 k€ every 3 months

Application Level Performance indicators	Execution Time Data Quality	Time to produce automated decisions Accuracy of new models with respect to internal CRF Aluminium Casting models	1 month ND	Near real time 20%
Platform level performance indicators	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization	30 k€	Platform costs

Table 31: Experiment #8 - Production process of aluminium casting

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
Business KPIs	Product quality	Quality control 1	67%	Increase of 2-6% (objective=77%)
		Quality control 2	27%	Decrease of 1-4% (objective=22 %)
		Quality control 3	6%	Decrease of 0-2% (objective=1%)
Application Level Performance indicators	Execution Time	Time to produce automated decisions	1 month	A few seconds
	Data Quality	Accuracy of new models with respect to internal CRF Aluminium Casting models	ND	Accuracy of models is better than the current models (6% increase)
Platform level performance indicators	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization	30 k€	Platform costs

4.6.4 Generic use case experiments

Table 32 and Table 33 provide an overview of the metrics and associated expected values with respect to the two generic I-BiDaaS experiments.

Table 32: Experiment #9 - Experiment for end-to-end I-BiDaaS platform in self-service mode

	Indicator	Metric	Expected I-BiDaaS value
Business KPIs	Operability (real business resting possibilities) (Q4)	Evaluator's score on scale 1-5	A score greater or equal to 3

	Innovation (bring innovative features) (Q4)	Evaluator's score on scale 1-5	A score greater or equal to 3
	Compliance (assurance compliance with the current security and privacy regulations) (Q4)	Evaluator's score on scale 1-5	A score greater or equal to 3
	Cost reduction (to what extent I-BiDaaS solution could be cost-effective compared to current solutions) (Q4)	Evaluator's score on scale 1-5	A score greater or equal to 3
Application Level Performance indicators	User guidance and usability(Q1/Q2)	Evaluator's score on scale 1-5	A score greater or equal to 3
	Information output(results) (Q1/Q2)	Evaluator's score on scale 1-5	A score greater or equal to 3
Platform level performance indicators	Correctness (Q3)	Number of crashes or malfunctions while using the platform	Zero
	Effectiveness (Q3)	Number of tasks completed / Total number of tasks	100

Table 33: Experiment #10 - Experiment for end-to-end I-BiDaaS platform in expert mode

	Indicator	Metric	Expected I-BiDaaS value
Business KPIs	Time efficiency	Productivity improvement (Decrease, no change, slight increase, moderate increase, high increase)	Subjective assessment of the experimental subject with regards to the I-BiDaaS solution (Decrease, no change, slight increase, moderate increase, high increase w.r.t. previous experience)
	Service quality	Correspondence of I-BiDaaS platform to user needs (scale 1-5)	A score greater or equal to 3
	Cost/Effort reduction (because of not having to maintain or setup a PyCOMPSs environment)	Evaluator's score on scale 1-5	A score greater or equal to 3
Application Level Performance indicators	Usability with respect to each element 1-8 in the workflow	Evaluator's score on scale 1-5 for each workflow element	A score greater or equal to 3
	Coding effort	Lines of code to develop a new algorithm based on code template	33% of the number of lines in the code template
Platform level performance indicators	Correctness	Number of code templates completed without problems (expressed as percentage)	100%

	Effectiveness	Number of tasks completed / Total number of tasks	100%
--	---------------	---	------

4.6.5 High-level non-functional requirements evaluation

All experiments have been tested considering the evaluation of usability, operability, robustness, innovation, compliance, privacy awareness and cost of the I-BiDaaS solution, based on the user questionnaires filled by the data providers (see D6.3 [2])

Specifically:

- Usability: The usage of real and valuable datasets from different industrial sectors allowed to test the usability of the I-BiDaaS solution with three different mode, developed for expert and non-expert users. The data anonymization was used in this process to ensure data privacy as well as the preservation of sensitive information while delivering the desired data analytics and meta-knowledge. The advanced visualisation tools provided a simple and intuitive design, by showing the qualities of the underlying data analytics and meta-information and delivered to the platform the expected added value.
- Operability: The I-BiDaaS solution can be integrated in a real business setting, taking into account of the specific internal requirements that have to be evaluated on a case-by-case basis. The different use cases tested within I-BiDaaS platform show that it is not only high performing but also scalable, by guarantying security and privacy.
- Robustness: The same use cases have been tested with several datasets and the results were valid in all the tests. Furthermore, several iterations were performed and they provided the same results.
- Innovation: The main innovation lies in the capacity to easily empower both non-expert and expert big data practitioners, belonged to a very diverse use case landscape, which is supported by a high level algorithmic solutions. I-BiDaaS platform provides not only the end-user mode but also the expert mode that enables the data analyst to directly prepare the dataset inside the same platform in the cloud and does so by leveraging advanced visualisation approaches and dashboards that harness the power of multiple heterogeneous sources and Big Data analytics. This facilitates the ability to take, to understand, to process, to visualize data and to extract value from them.
- Compliance: The relevant security and privacy regulations are put in place internally, during the selection and pre-processing stages when data are anonymised and tokenised. In this way, data providers manage internally information and data are made readily available without requiring further security or anonymization steps to be implemented by the I-BiDaaS platform.
- Privacy awareness: All relevant security and privacy requirements pertaining to the in-house access of proprietary data were met and the necessary practices were applied (e.g., data anonymization, data aggregation, encryption, etc.).
- Cost: The I-BiDaaS platform allows to reduce infrastructure cost and personnel cost. In the first case, it allows to obtain improved results and state-of-the-art performance using less hardware resources, thus cutting down on unnecessary and costly investments and avoiding the maintenance of expensive infrastructure. In the second case, it replaces certain manual, labour-intensive and costly practices with automated, efficient and scalable technologies.

5 Impact Analysis

I-BiDaaS is delivering a full array of big data business analytics solutions for real and synthetic data for companies in the domains of telecommunication, finance and manufacturing that are more accessible, cost-effective and employee-empowering than existing solutions, which gives companies the opportunity to deploy Big Data Self-Service solutions across the organisation, from consumer-facing employees with little IT experience or expertise to top management, and helps companies to optimize decision-making at the tactical, operational and strategic levels.

To ensure that the I-BiDaaS project meets its ambitious objectives, and achieve expected impacts, it is using a three-stage impact assessment model that realises the project's business case, monitors progress, raises any issues and helps inform operational decisions. The impact assessments occur at the: a) Project Level to ensure Project Partners deliver the required outputs to test the business cases; b) Pilot Level with involved Local and National Stakeholders to produce outcomes that test and refine the value proposition and improve the business case for I-BiDaaS; and c) European Level encompassing wider society to aggregate and spread social and economic benefits that result from the business case.

In this section, the results of the Experimentation & Evaluation Phase (M19-M32) of the project, along with an analysis with respect to the expected project level innovation and achievements, are discussed. Moreover, the implemented or prospective activities aiming to demonstrate the I-BiDaaS solution and involve external users in the evaluation process, are described.

5.1 Progress Report

After the end of M18, and the successful completion of the Innovation phase, the I-BiDaaS project entered into the Experimentation and Evaluation phase where all functionalities developed during the previous period are implemented on 8 (eight) real-life industrial scenarios in the I-BiDaaS targeted domains of telecommunication, finance and manufacturing. During the Experimentation and Evaluation phase, all the innovation development results accomplished in the previous phase under the technical WPs of the project (WP2-WP5) have been integrated, realising the 2nd version of the I-BiDaaS platform as an integrated framework that enables effective extraction of meaningful knowledge from integrating very large datasets from heterogeneous and multiple domains and more effective and scalable data analytics and real-time complex event processing to support low-level employees and decision makers with advanced visualization capabilities.

The Experimentation phase is being supported by various events organized by I-BiDaaS, such as the CAIXA's Workshop, the Big Data Pilot Demo Days series of webinars where 3 (three dedicated webinars were organized under BDV PPP Summit 2020. The I-BiDaaS Solution has also been demonstrated in various events such as the BDV PPP Summit 2019 in Riga and the EBDVF 2019 in Helsinki. Therein, the experimentation is supported through the involvement of external entities in the evaluation of the I-BiDaaS solution and feedback collection from these entities. In all the aforementioned events, the main goal is to promote I-BiDaaS tools and technologies to software developers, big data experts, data analysts, decision makers, non-IT end users, etc., and receive valuable feedback for further improvement. More events are planned to be executed until the end of the project (e.g., TID's Hackathon and a big event at the end of the project to celebrate all I-BiDaaS achievements). The output of this phase will feed the Consolidation Phase that is following.

I-BiDaaS, from the very beginning, has defined specific, ambitious objectives while increasing the impact in the research community, in Data Market and the Big Data Economy and contributing to innovation capacity.

I-BiDaaS contribution towards each of the expected impacts mentioned in the work programme: I-BiDaaS achievements are closely monitored through the Key Performance Indicators (KPIs) defined by the consortium. In the following tables (Table 34 and Table 35), the current status of the KPIs is reported.

Table 34: Progress with regards to I-BiDaaS KPIs

What the call states		
KPI-RI-1	Release of I-BiDaaS framework and tools under an open-source non-viral license.	5
KPI-RI-2	Increased speed of data analysis and throughput (compared to industrial-based benchmarks) by more than 10%.	Work in Progress
KPI-RI-3	Increase in the direct access of big data analytics tools by more than 30%.	100%
KPI-RI-4	Define at least 2 standards related to Big Data Analytics and uptake at least 5.	1
KPI-RI-5	Influence at least 4 formal specifications of standards.	2
KPI-RI-6	Implementation of 3 data practitioners' demonstrators validating at least 80% of tools.	8
Impact & Exploitation KPIs		
KPI-IE-1	At least 3 I-BiDaaS tools reach market readiness level at the end of the project.	1
KPI-IE-2	At least 4 standalone tools and methods delivered.	4
KPI-IE-3	At least 6 third-party collaborations to be established for further applicability verification.	>6
KPI-IE-4	At least 3 experiments demonstrating the tools' applicability within I-BiDaaS.	10
KPI-IE-5	Increased programmability for users by at least 30% compared to today, verified on at least 1 practitioner.	~57% less LoCs using Hecuba & COMPSs for distributed programming & DB access
KPI-IE-6	Reduction of practitioners LOCs (lines of code) by 50% due to the ability to transform a sequential application into a parallel and distributed one.	~50%
KPI-IE-7	At least 1500 downloads of the tools through the project.	Work in Progress
Dissemination & Communication KPIs		
KPI-DC-1	At least 500 downloads for public deliverables, prototypes, promotional material.	1144 Direct Downloads
KPI-DC-2	At least 10 publications.	3 Journal, 17 Conference and Workshop Papers, 2 Posters
KPI-DC-3	At least 3 conferences or workshop participations per year.	16 Conferences, 7 Workshops

KPI-DC-4	At least 33% of conference and journal papers have an impact factor or ERA classification.	64%
KPI-DC-5	At least 33% gold open-access journal articles.	33.3%
KPI-DC-6	At least 2 active participations to a standardization body.	2
KPI-DC-7	At least 2 standards that are used and improved within I-BiDaaS.	Work in Progress
KPI-DC-8	At least 3 workshops or special events.	6
KPI-DC-9	At least 3 collaborations with projects in H2020.	3
KPI-DC-10	At least 4 participations to collaborative initiatives.	2

Table 35: Explanation of the status and the next steps of the I-BiDaaS KPIs

What the call states	
KPI-RI-1	One of the goals of the I-BiDaaS consortium is to contribute to the Open Source communities to provide benefit to the European community. COMPs and Hecuba are provided under an open-source license by BSC and also a pool of ML algorithms based on structured (non) convex optimization by UNSPMF that is being enriched continuously. The selected stack used to support the cloud service management of resources includes technology enablers and tools which are realized under open-source non-viral licenses. Additionally, license analysis has been made within the consortium in order to identify technology candidates that in the future can substitute the modules of the platform that has not been released as open-source.
KPI-RI-2	Since we use Streaming Analytics and in-memory techniques, we believe that data analysis and throughput is fast and will have a high user acceptance rate. We still need to find a suitable benchmark and other comparable platforms to which we can compare our performance.
KPI-RI-3	The I-BiDaaS data providers (thanks to IBM's TDF) created new synthetic data sets that were not available before. In addition, CAIXA, TID and CRF created within I-BiDaaS new anonymised data sets. These synthetic and real data sets are analysed within the I-BiDaaS platform or harnessing the I-BiDaaS partners expertise. In this way, the I-BiDaaS data providers are given access to tools and technologies which were not available to them before. In addition, considering the CAIXA case, by uploading the anonymized and synthetic data to the I-BiDaaS platform (ATOS cloud), CAIXA makes Big Data experimentation more agile, as experimenting in house may be time and resource consuming due to internal bank processes.
KPI-RI-4	Contributing to one standard via BDVA. As of August 2017, BDVA has an official liaison with ISO IEC JTC1 WG9 Big Data Standards group merging into JTC 1/SC 42, Artificial Intelligence that is developing the Big Data Reference Architecture for ISO.
KPI-RI-5	All I-BiDaaS data providers (CAIXA, CRF, TID) contributed feedback by completing the questionnaire for DataBench project. The I-BiDaaS components closely match the functional components of the Big Data Interoperability Framework of NIST. The I-BiDaaS architecture addresses most of the horizontal concerns identified in the BDVA SRIA, in the context of the BDV reference model. Referencing these organizations is influencing.
KPI-RI-6	8 use case demonstrators (three from CAIXA, two from CRF and three from TID) are already deployed and tested in the I-BiDaaS platform (and thus validating all the tools provided and integrated in I-BiDaaS). All demonstrators have been deployed according to the plan.
Impact & Exploitation KPIs	

KPI-IE-1	Qbeast by BSC has reached market readiness level and have been commercialized in 2020. Qbeast is a spin-off company originated at Barcelona Supercomputing Center (BSC), which focuses on the analysis of Big Data with approximate analytics, reducing both the time and cost of obtaining knowledge in big collections of data. The Qbeast framework has been partially developed within I-BiDaaS.
KPI-IE-2	I-BiDaaS delivers several standalone tools. For example, IBM's TDF, AEGIS' Advanced visualization toolkit, BSC's Hecuba, and UNSPMF's machine learning algorithm implementations in COMPSSs. Each of the mentioned tools is either fully created within I-BiDaaS or is significantly improved within I-BiDaaS.
KPI-IE-3	Third-party collaborations are focusing on SMEs and have already been established. During the course of the project, the I-BiDaaS Consortium organised several events such as Workshops, Info Days, Hackathon, Webinars and participated to even more events (BDV PPP Summit 2019, EBDVF 2019, BDV PPP Summit 2020) aiming to demonstrate the I-BiDaaS Solution and its application to the three different sectors (finance, telecommunication, manufacturing). During all these events, I-BiDaaS managed to engage different SMEs that are actively following all the I-BiDaaS updates. Moreover, the dissemination team of I-BiDaaS has already initiated targeted efforts to engage SMEs communities such as Engagement of Serbian Data Science community, engagement of Praxi Network and engagement of SME Community by ENPC as reported in D7.5 [22]. I-BiDaaS will continue the same approach until the end of the project so as to boost the project's efforts to seek for even more collaborations.
KPI-IE-4	<p>So far, 8 real-life experiments have been performed that demonstrate the applicability of the I-BiDaaS solution to address 8 real problems in the telecommunication, financial and manufacturing sectors. These experiments reflect the specific requirements of the project industrial partners and correspond to the 'Co-Develop' mode of operation of the I-BiDaaS platform.</p> <p>In addition, two generic experiments have been defined aiming to demonstrate how the I-BiDaaS solution can be applied generally, reflecting the requirements of generic user categories, corresponding to the 'Expert' and 'Self-Service' mode of operation of the I-BiDaaS platform. The operation of these experiments is currently in progress.</p>
KPI-IE-5	<p>The programmability is highly improved as users do not have to learn new programming paradigms or to change the way to manage the data. When users require high-level data manipulation, they can use our algorithm integrated into Dislib, so that they can define in a declarative way how to analyse data while our framework takes care of distribute the execution across multiple nodes. On the other hand, when users need more flexibility manipulating they can develop their code with a familiar imperative approach, without having to worry about parallelization, fault-tolerance, persistence, and data access, as COMPSSs and Hecuba take care of it.</p> <p>Taking as a simple example</p> <p>https://github.com/ibidaas/knowledge_repository/tree/master/tools_technologies/compare_LOC</p> <p>and programmed it with / without COMPSSs + Hecuba. If you account for meaningful code only (skipping definitions) it is 73 lines of code vs 171, which is a ~57% reduction.</p>
KPI-IE-6	As an example of how this KPI is reached, we compare an MPI ³⁸ and a COMPSS implementation ³⁹ for the ADMM-Lasso algorithm for sparse regression. The MPI (C-based) implementation has 216 lines of code, while the COMPSSs (Python-based) implementation has 109 lines, which yields approximately a 50% reduction ⁴⁰ .

³⁸ <https://web.stanford.edu/~boyd/papers/admm/mpi/>

³⁹ https://github.com/ibidaas/knowledge_repository/blob/master/tools_technologies/sources/batch_processing/uns_pmf/distributed_Lasso_ADMM.py

⁴⁰ The programming languages are different (C-based MPI versus Python based COMPSSs), but the comparison is reasonable, as a lower-level language (C) yields more code for a tighter resource control, while with COMPSSs we have fewer lines of code at a price of moderately reduced performance; see Table 21, Advanced ML submodule.

KPI-IE-7	All the I-BiDaaS tools (either open-source or proprietary) are listed at the Tools section of the I-BiDaaS website. The open-source code is available at GitHub. Since the number of downloads/clones of the git hub repo is not provided by GitHub, we are counting the popularity of the Tools section (pressed links to the knowledge database and pressed links to the proprietary tools of the project). The number of events (pressed links) is expected to be reach the KPI threshold due to the release of the 2 nd and final version of the tools developed within the project and also due to the continuous dissemination efforts until the end of the project.
Dissemination & Communication KPIs	
KPI-DC-1	Although this KPI is covered, the project will continue the dissemination efforts to increase the downloads of its material.
KPI-DC-2	I-BiDaaS is delivering a solid publications list, including 3 Journal papers, 15 Conference/Workshop papers, and 2 Conference/Workshop Posters. The detailed list for 2018 publications is reported in D7.3 [21] and for 2019 in D7.5 [22]. I-BiDaaS publications can be found on the project's website ⁴¹ and also in Zenodo ⁴² and OpenAIRE ⁴³ .
KPI-DC-3	<p>During the reporting period, I-BiDaaS partners attended 11 Conferences & 5 Workshops</p> <ul style="list-style-type: none"> • 2018: 7 Conferences & 4 Workshops • 2019: 9 Conferences & 1 Workshop • 2020: 3 Conferences & 1 Workshop <p>The detailed list of conferences & workshops attended by I-BiDaaS partners can be found at D7.3 [21] for 2018 and D7.5 [22] for 2019. For 2020, the list of the events will be reported at D7.7 (to be submitted at M36).</p> <p>I-BiDaaS Consortium invests in events targeted at industry and academia to showcase I-BiDaaS vision, impact and results, and to create an active community for the project that will significantly enhance its entrance to the market.</p>
KPI-DC-4	For this reporting period, we have achieved 64% of the publications to have an impact factor or ERA classification.
KPI-DC-5	<p>During the reporting period, 3 (three) journal articles have been accepted and published.</p> <ol style="list-style-type: none"> 1. Gold Open Access - Sahu, A.K., Jakovetic, D., Bajovic, D. and Kar, S., 2018. Communication efficient distributed weighted non-linear least-squares estimation. EURASIP Journal on Advances in Signal Processing, 2018(1), p.66. 2. Green Open Access - Jerinkić, N.K., Jakovetić, D., Krejić, N. and Bajović, D., 2019. Distributed second-order methods with increasing number of working nodes. IEEE Transactions on Automatic Control. 3. Green Open Access - Jakovetić, D., Krejić, N. and Jerinkić, N.K., 2019. Exact spectral-like gradient method for distributed optimization. Computational Optimization and Applications. <p>Thus, I-BiDaaS consortium has achieved 33,3% gold open access to the journal articles linked to I-BiDaaS scientific results score higher than the expected for this KPI.</p>
KPI-DC-6	<p>Participation in BDVA which is driving big data standardization and interoperability priorities and is connected with Big Data Standards related to Big Data PPP projects</p> <p>Participation and active collaboration with DataBench, who is designing performance benchmarking processes for Big Data. DataBench is expected to set the standards and benchmarks for the emerging Big Data ecosystem.</p>

⁴¹ <http://www.ibidaas.eu>⁴² <https://zenodo.org>⁴³ https://explore.openaire.eu/search/project?projectId=corda_h2020:652e6b81a75292294cdd34ff5a806573

KPI-DC-7	<p>WP1: NIST and BDVA big data reference models</p> <p>WP2: MQTT (Message Queuing Telemetry Transport), JSON (Javascript Object Notation)</p> <p>WP4:</p> <ul style="list-style-type: none"> • SAG's Universal Messaging supports the following standards: • JMS – Java Message Service • AMQP – Advanced Message Queuing Protocol • MQTT – Message Queuing Telemetry Transport (an ISO standard) • SAG's Apama supports the following standards: • JDBC – Java Database Connectivity • PMML – Predictive Model Markup Language <p>WP5: ATOS manages the Cloud environment in accordance with ISO 9001 (quality), ISO14001 (environmental), and ISO 27001 (security) standards.</p> <p>Standards and Bodies of Knowledge: BABOK, CMMI, IEEE standards, ISO 9001, ISO/IEC standards, PMBOK, SWEBOK, ITIL</p>
KPI-DC-8	<p>Six (6) special events have been organized in the context of I-BiDaaS:</p> <ol style="list-style-type: none"> 1. I-BiDaaS Info Day - Workshop on Big Data Analytics⁴⁴⁴⁵: January 22, 2019, Faculty of Sciences of University of Novi Sad, Serbia 2. CRF's hackathon at Campus Melfi: June 18-19, 2019, Campus Melfi, Italy. 3. Satellite Promotional Event at BDV PPP Summit in Riga: June 26-28, 2019, Riga, Latvia 4. European Big Data Value Forum 2019: October 14-16, 2019, Helsinki, Finland 5. CAIXA Virtual Workshop: June 22, 2020 6. BDV PPP Virtual Summit 2020 - Big Data Pilot Demo Days series of webinars (Collaboration between I-BiDaaS and BigDataStack) May 21 – July 16, 2020 <ul style="list-style-type: none"> a. I-BiDaaS Application to the Financial Sector, May 21, 2020⁴⁶ b. I-BiDaaS Application to the Telecommunication Sector, June 25, 2020⁴⁷ c. I-BiDaaS Application to the Manufacturing Sector, July 9, 2020⁴⁸
KPI-DC-9	<p>Three (3) collaborations with H2020 projects have been initiated:</p> <ol style="list-style-type: none"> 1. TOREADOR - TrustwOrthy model-awaRE Analytics Data platform (GA #688787)⁴⁹ 2. DataBench - Evidence-Based Big Data Benchmarking to Improve Business Performance (GA #780966)⁵⁰ 3. BigDataStack – Holistic Stack for BigData Applications and Operations (GA # 779747)⁵¹
KPI-DC-10	<p>I-BiDaaS is actively participating in most (if not all) BDVA activities. Moreover, I-BiDaaS participates in ENISA⁵² activities since I-BiDaaS Coordinator Dr. Sotiris Ioannidis is a PSG member of ENISA.</p>

⁴⁴ <http://www.ibidaas.eu/events/I-BiDaaS-Info-Day-Workshop-on-Big-Data-Analytics>

⁴⁵ <https://ibidaas.eu/blog/I-BiDaaS-Info-Day-Workshop>

⁴⁶ <http://www.ibidaas.eu/events/Big-Data-Pilot-Demo-Days%3A-I-BiDaaS-Application-to-the-Financial-Sector>

⁴⁷ <http://www.ibidaas.eu/events/I-BiDaaS-Application-to-the-Telecommunication-Sector-Webinar>

⁴⁸ <http://www.ibidaas.eu/events/I-BiDaaS-Application-to-the-Manufacturing-Sector-Webinar>

⁴⁹ <http://www.toreador-project.eu/>

⁵⁰ <https://www.databench.eu>

⁵¹ <https://bigdastack.eu>

⁵² <https://www.enisa.europa.eu>

Moreover, the evolution of the different modules of the I-BiDaas platform is progressing as expected as well, as it can be seen in the following table:

Table 36: I-BiDaas Tools & Technologies

Tool	Partner	Initial TRL	Current TRL	Expected TRL
GPU accelerator technology	FORTH	3	5	5
COMPSSs (Programming Model and Runtime)	BSC	8	9	9
Hecuba	BSC	5	6	6
QBeast	BSC	5	7	6
Test Data Fabrication	IBM	6	6	7
Apama Streaming Analytics Platform	SAG	6	6	7
Universal Messaging	SAG	7	7	8
MashZone	SAG	4	4	5
Pool of ML algorithms based on structured (non)convex optimization	UNSPMF	3	5	5
Orchestrator	ITML	-	4	5
Advanced visualization and monitoring	AEGIS	4	5	6
Interface services/I-BiDaas application	AEGIS	4	5	6
Resource management and orchestration module – Adaptation Engine Submodule	ATOS	2	4	5
Resource management and orchestration module – Cloudify Cloud Orchestrator Submodule	ATOS	9	9	9
Resource management and orchestration module – Private Cloud based on Openstack Submodule	ATOS	9	9	9

Impact in research community and contribution to innovation capacity: The I-BiDaas project has achieved a significant impact in the research community with 3 journal publications, 17 conference papers and 2 poster publications, with more than 60% of these publications having an impact factor or ERA classification. Moreover, an important achievement for the I-BiDaas project is considered the acceptance of 5 (five) Innovations developed under I-BiDaas from EU Innovation Radar as Excellent Innovations. The Innovation Radar is a European Commission initiative to identify high potential innovations and innovators in EU-funded research and innovation framework programmes. The full list of the accepted Innovations is depicted in Table 37:

Table 37: List of I-BiDaas Excellent Innovations as accepted from EU Innovation Radar⁵³

Innovation	Key Innovators	Market Maturity	Innovation Topic
Multidimensional Storage with Efficient Sampling (MuSES)⁵⁴	BSC	Market Ready	Deep Tech
Advanced Visualization Toolkit (AVT) supporting scalable data visualisation⁵⁵	AEGIS	Business Ready	Deep Tech

⁵³ <https://www.innорадар.eu/resultbykeyword/I-BiDaas>

⁵⁴ <https://www.innорадар.eu/innovation/35296>

⁵⁵ <https://www.innорадар.eu/innovation/35294>

ADMM Machine Learning Algorithms⁵⁶	BSC UNSPMF	Exploring	Deep Tech
Parallelization of constraint satisfaction problems⁵⁷	IBM	Exploring	Deep Tech
Specification of an end-to-end Big Data as-a-self-service platform⁵⁸	UNIMAN ATOS UNSPMF	Exploring	Deep Tech

Impact in Data Market and the Big Data Economy: I-BiDaaS is developing a Big Data as a Self-Service Solution to provide a significant boost to the finance, manufacturing and telecommunication sector. For the financial and telecommunication sectors, I-BiDaaS has offered its tools and services for the development of 6 (six) different use cases (3 use cases for CAIXA and 3 use cases for TID) making possible for CAIXA and TID to exploit their big data efficiently and therefore increase their market share and services provided to their customers. For the manufacturing sector, I-BiDaaS has offered its tools and services for the development of two different use cases making possible for CRF for even easier and massive big data exploitation.

Contribution to standards and international initiatives: I-BiDaaS is fully aligned with all PPP BDVA activities like the European Big Data Value Strategic Research and Innovation Agenda (BDVA SRIA), aiming to ensure the viability of the I-BiDaaS solution. The I-BiDaaS architecture addresses most of the horizontal and vertical concerns identified in the BDVA SRIA, in the context of the BDV reference model. Thus, it provides a BDVA reference model-compliant end-to-end Big Data solution. Moreover, I-BiDaaS has initiated a collaboration with DataBench project both from a technical and a business perspective. I-BiDaaS data providers contributed to the DataBench survey in order to develop in-depth case studies of business KPIs of Big Data. Currently, the technical and business teams of both projects are collaborating to explore how to best exploit the DataBench Toolbox in the context of I-BiDaaS

5.2 Feedback from external stakeholders

In this section, we present the external stakeholders' feedback received during the project period M18-M32; for earlier results, the reader may refer to D6.1 [19] and D6.3 [2]. In particular, we discuss the following important events at which the external feedback was collected: Big Data Pilot Demo Days, CAIXA workshop, and the webinar 'Virtual BenchLearning – Assessing the Performance and Impact of Big Data, Analytics and AI' organized by the DataBench project.

Big Data Pilot Demo Days. Hosted by BDV PPP and in the scope of BDV PPP Summit 2020⁵⁹, the EU projects BigDataStack, I-BiDaaS, Track & Know, and Policy Cloud have jointly organized a series of webinars and online demonstrations of innovative Big Data Technologies. Within this series of events, I-BiDaaS has organized three webinars, namely I-BiDaaS Application to the Financial Sector, held on May 21, 2020; I-BiDaaS Application to the Telecommunication Sector; held on June 25, 2020; and I-BiDaaS Application to the

⁵⁶ <https://www.innoradar.eu/innovation/35298>

⁵⁷ <https://www.innoradar.eu/innovation/35295>

⁵⁸ <https://www.innoradar.eu/innovation/35293>

⁵⁹ <https://www.big-data-value.eu/bdvpsummit-2020/>

Manufacturing Sector, held on July 9, 2020. The main goal of the webinars was to demonstrate in a step by step fashion the I-BiDaaS solution in the three sectors and receive feedback from the participants. At each of the three events, the audience was asked to respond to four short questions, aiming to investigate the background of the attendees and adjust the nature of the webinars accordingly. The questions were: 1) to which stakeholder type they belong; 2) whether they work with Big Data; 3) if they are interested in Big Data technologies to optimize customer experience; and 4) what is the main barrier from preventing the Big Data analytics technologies in their organization. The results of the questions for each of the three events are shown in Figure 45.

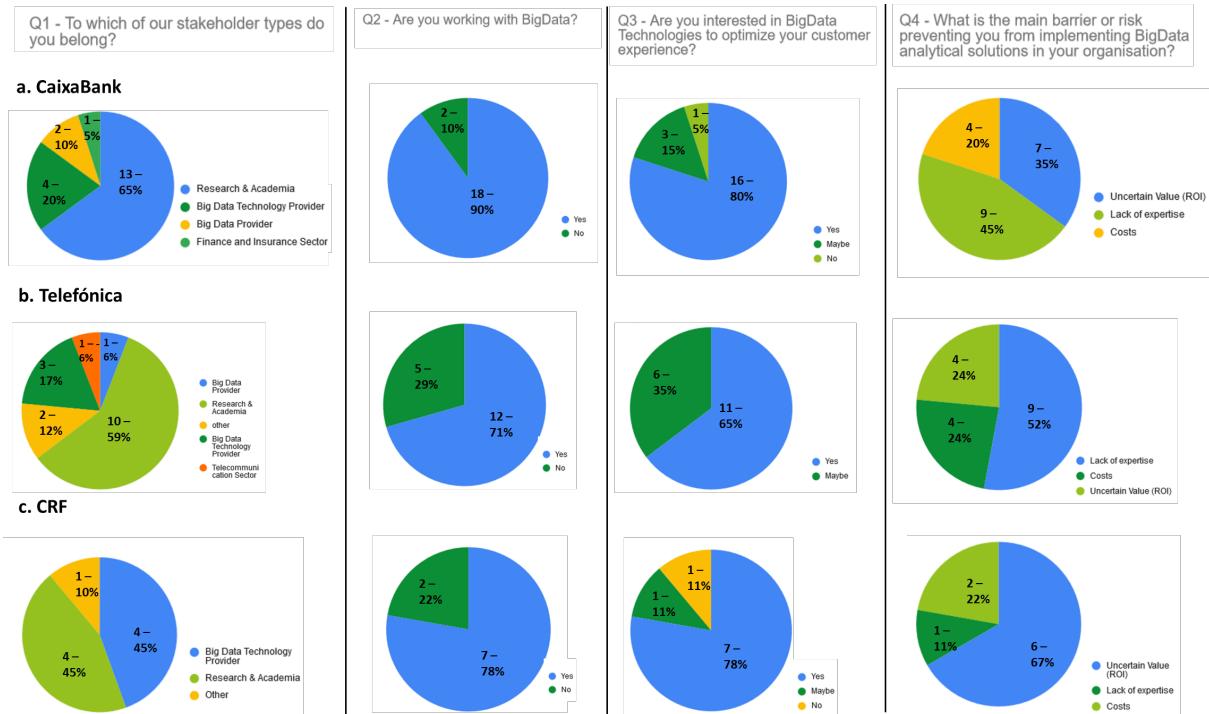


Figure 45: General questions webinar results: a. CaixaBank; b. Telefonica; c. CRF

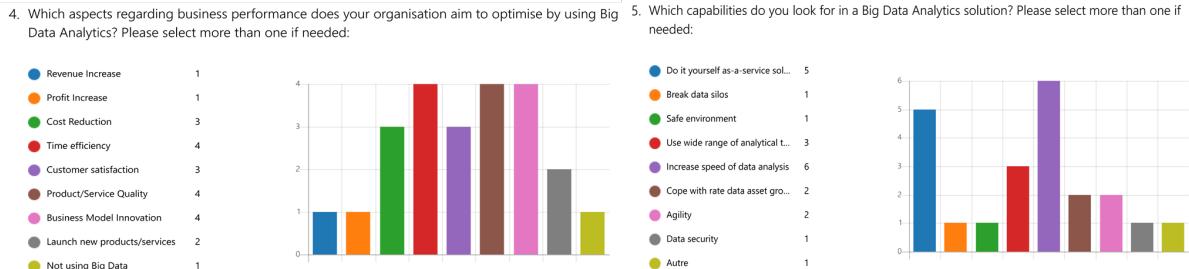
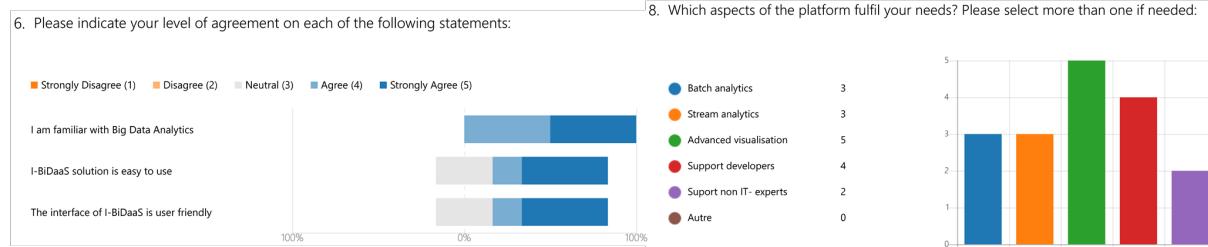
In addition, after each of the events, a more detailed questionnaire designed using Microsoft forms has been provided to the attendees. The respective questionnaires (CaixaBank⁶⁰, Telefonica⁶¹, CRF⁶²) can be found in the Appendix. The presentation of an example of the results is shown in Figure 46. More details about the three webinars can be found at the I-BiDaaS website, in the news & events section⁶³.

⁶⁰ <https://bit.ly/2Zl19IJ>

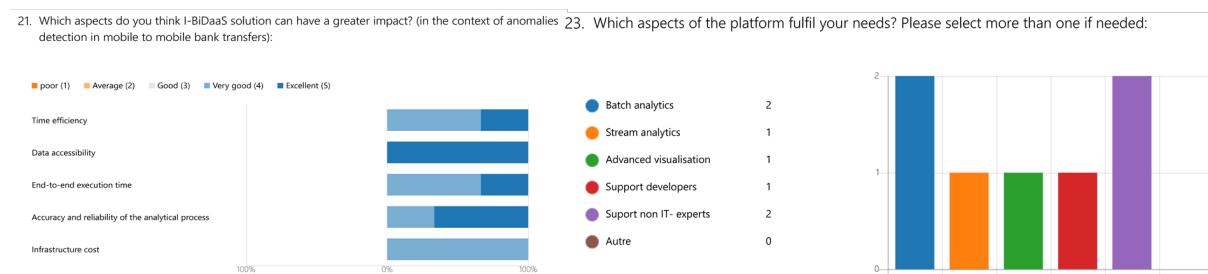
⁶¹ <https://bit.ly/3dtXOdX>

⁶² <https://bit.ly/2O8YV8D>

⁶³ <https://www.ibidaas.eu/events>

a. Telefonica**b. CRF****Figure 46: Webinar results specific questionnaire: a. Telefonica; b. CRF**

CAIXA Workshop. CAIXA organized an online I-BiDaaS workshop on June 22, 2020. The workshop's participants were professionals in the Big Data, data science, and related domains that are employed at CAIXA (but are not part of the I-BiDaaS team) or collaborate with CAIXA. The workshop included several presentations, namely 1) a presentation of the requirements in the financial/banking sector relevant to I-BiDaaS; 2) the I-BiDaaS solution (architecture) with an emphasis on how it addresses the banking sector requirements; and 3) the CAIXA use cases provided to I-BiDaaS followed by step-by-step demonstrations of the I-BiDaaS solutions to the respective use cases. In addition, the workshop included a hands-on session where the participants were able to access the I-BiDaaS online platform and experiment with the platform in the self-service mode. The workshop participants were asked to fill out a detailed questionnaire about the I-BiDaaS platform, provided in the Appendix. An example of the results of the questionnaire is shown in Figure 47 below.

CaixaBank**Figure 47: CaixaBank Workshop results**

Virtual BenchLearning - Assessing the Performance and Impact of Big Data, Analytics and AI. On July 7, 2020, I-BiDaaS participated in this webinar organized by the DataBench project. The webinar described a framework and tools to assess the performance and impact of Big Data and AI technologies by providing real insights coming from DataBench. I-BiDaaS participated at the webinar through a presentation of the current I-BiDaaS benchmarking approach, landscape and needs, both from the technological and business perspectives. The main goal of our participation is to explore, during the final 6-months period of the project, the

possibilities to harness collaboration and benchmarking tools provided by the DataBench project.

5.3 Exploitation and potential commercialization

According to Statista's IT Market Model, spending in the global IT services market will reach an 853 billion U.S. dollars mark by 2021, up from 737 billion U.S. dollars in 2017. Profitability and cost reduction are some of the expected impacts from I-BiDaaS, providing companies with the competitive advantage they need towards a thriving data-driven EU economy. The nature of the project requires designing new and innovative businesses that often include complex interconnections and interoperable dimensions.

The use of the Dynamic Business Modelling (DBM) tool and the exploitation workshop (see D7.6 [23]) enabled the definition of the exploitation strategy for each participating partner, as well as for the whole consortium, maximizing the exploitation opportunities for individual partners and the sustainability of the tools in the long-term, beyond the lifespan of the project. Moreover, individual exploitation plans enabled the identification of potential I-BiDaaS products linked to actual market needs, given their capability to address different stakeholders in the market.

Five different joint business models were developed to adapt each of the solutions and business processes to the targeted markets and clients, including Non-IT SMEs, Large companies, Academics, and Data harvesting companies. The business models will be re-analysed in the last months of the project for the design and proposition of a sound business plan to ensure the long-term sustainability and potential commercial viability of the solution. Furthermore, the joint exploitation plan will also include a detailed profiling of partners (including academics) to develop a competency profile for the whole consortium and allow the export of ‘the I-BiDaaS innovation ecosystem’ to third-party companies.

The design of the I-BiDaaS solution allows to decouple a given tool from the rest, making the platform flexible and modular, and these factors will impact the pricing model. Therefore, a methodology for the pricing model was designed, enabling the identification of a suitable licensing strategy and possible collaboration that will allow the establishment of the agreement for partnerships. Business planning activities will be analysed for the elaboration of any Intellectual Property Rights (IPR) aspects and revenue sharing models, including an analysis of the potential revenue streams to validate the Return on Investment (RoI) plans.

Finally, to achieve long-term sustainability, I-BiDaaS will leverage the opportunity and ambition of the EU to become a global leader in the acceleration towards digital transformation as well as the purpose of Europe to become a circular industry, by the incorporation of circular business models.

6 Conclusion

This deliverable reports on the experimentation phases of the industrial experiments carried out within the I-BiDaaS project. It provides a detailed description of each experiment in terms of the dataset implemented and the experimental workflow. Pilot demos have been developed for each use case and they can be accessed via the I-BiDaaS platform, integrated for expert and non-expert users. Furthermore, for each real industrial sector, end-users have been described and it has shown how to easily utilise I-BiDaaS solution.

All of work presented is aligned with the experimental protocol described in D6.3 [2] and revised during the implementation of each experiment to assure that the designed experiments validate both business and technical requirements in the I-BiDaaS platform and associated technology characteristics.

In the final section of this deliverable, we present the benefits of the participatory evaluation that involved the consortium and external stakeholders in evaluating key results and what constitutes success. Furthermore, for each experiment, the impact analysis has been carried out, by focusing on usability, operability, innovation, robustness, privacy awareness and cost of the I-BiDaaS solution. The final results will be reported in the next deliverable D6.5 '*Assessment report and impact analysis*' (M36).

7 References

- [1] I-BiDaS Consortium, Deliverable D6.2: Experiments implementation-initial version, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [2] I-BiDaS Consortium, Deliverable D6.3: Evaluation report (final version), Confidential Report
- [3] I-BiDaS Consortium, Deliverable D1.3: Positioning of I-BiDaS, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [4] I-BiDaS Consortium, Deliverable D2.1: Data assets and formats, Confidential Report
- [5] I-BiDaS Consortium, Deliverable D2.5: The Data Fabrication Platform (DFP, final version), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [6] I-BiDaS Consortium, Deliverable D3.3: Batch Processing Analytics module implementation (final report), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [7] I-BiDaS Consortium, Deliverable D4.3: Streaming analytics and predictions, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [8] I-BiDaS Consortium, Deliverable D2.6: Universal Messaging Bus (final version), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [9] MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.
- [10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Evangelos Simoudis, Jiawei Han, and Usama Fayyad (Eds.). AAAI Press 226-231.
- [11] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.
- [12] Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 1 (November 1987), 53-65. DOI=[http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7)
- [13] Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E., Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008, 10008.
- [14] Rainer Schnell, Tobias Bachteler and Jörg Reiher, Privacy-preserving record linkage using Bloom filters, BMC Medical Informatics and Decision Making volume 9, Article number: 41 (2009)

- [15] 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.
- [16] Karr, A., A. Oganian, J. Reiter, and M.-J. Woo (2006). New measures of data utility. Workshop Manuscripts of Data Confidentiality, A Working Group in National Defense and Homeland Security.
- [17] Woo, M.-J., J. P. Reiter, A. Oganian, and A. F. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1, 111–124.
- [18] I-BiDaas Consortium, Deliverable D3.2: Batch Processing Analytics module implementation as part of I-BiDaas solution, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [19] I-BiDaas Consortium, Deliverable D6.1: Evaluation report (interim version), Confidential Report
- [20] I-BiDaas Consortium, Deliverable D3.1: Batch Processing Analytics module implementation, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [21] I-BiDaas Consortium, Deliverable D7.3: Dissemination strategy and activities (first report), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [22] I-BiDaas Consortium, Deliverable D7.5: Dissemination strategy and activities (second report), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [23] I-BiDaas Consortium, Deliverable D7.6: Exploitation strategy and activities (second report), Confidential Report

8 Appendix

Questionnaire for CaixaBank, Telefonica and CRF Webinars

<p>1. What is your role in your organisation?</p> <input type="text" value="Sélectionnez votre réponse"/>	<p>5. Which capabilities do you look for in a Big Data Analytics solution? Please select more than one if needed:</p> <ul style="list-style-type: none"> <input type="checkbox"/> Do it yourself as-a-service solution <input type="checkbox"/> Break data silos <input type="checkbox"/> Safe environment <input type="checkbox"/> Use wide range of analytical tools <input type="checkbox"/> Increase speed of data analysis <input type="checkbox"/> Cope with rate data asset growth <input type="checkbox"/> Agility <input type="checkbox"/> Data security <input checked="" type="checkbox"/> Autre 																																										
<p>2. What is the sector of your organisation?</p> <input type="text" value="Sélectionnez votre réponse"/>	<p>6. Please indicate your level of agreement on each of the following statements:</p> <table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th></th> <th>Strongly Disagree (1)</th> <th>Disagree (2)</th> <th>Neutral (3)</th> <th>Agree (4)</th> <th>Strongly Agree (5)</th> </tr> </thead> <tbody> <tr> <td>I am familiar with Big Data Analytics</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>I-BiDaaS solution is easy to use</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>The interface of I-BiDaaS is user friendly</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> </tbody> </table>		Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	I am familiar with Big Data Analytics	<input type="radio"/>	I-BiDaaS solution is easy to use	<input type="radio"/>	The interface of I-BiDaaS is user friendly	<input type="radio"/>																														
	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)																																						
I am familiar with Big Data Analytics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
I-BiDaaS solution is easy to use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
The interface of I-BiDaaS is user friendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
<p>3. Please provide the name of your organisation</p> <input type="text" value="Entrez votre réponse"/>	<p>7. In which aspects do you think I-BiDaaS solution can have a greater impact?</p> <table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th></th> <th>poor (1)</th> <th>Average (2)</th> <th>Good (3)</th> <th>Very good (4)</th> <th>Excellent (5)</th> </tr> </thead> <tbody> <tr> <td>Time efficiency</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Data accessibility</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>End-to-end execution time</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Accuracy and reliability of the analytical process</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Infrastructure cost</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Processing costs</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> </tbody> </table>		poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)	Time efficiency	<input type="radio"/>	Data accessibility	<input type="radio"/>	End-to-end execution time	<input type="radio"/>	Accuracy and reliability of the analytical process	<input type="radio"/>	Infrastructure cost	<input type="radio"/>	Processing costs	<input type="radio"/>																								
	poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)																																						
Time efficiency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Data accessibility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
End-to-end execution time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Accuracy and reliability of the analytical process	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Infrastructure cost	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Processing costs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
<p>8. Which aspects of the platform fulfil your needs? Please select more than one if needed:</p> <ul style="list-style-type: none"> <input type="radio"/> Batch analytics <input type="radio"/> Stream analytics <input type="radio"/> Advanced visualisation <input type="radio"/> Support developers <input type="radio"/> Support non IT-experts <input checked="" type="radio"/> Autre 	<p>9. How satisfied are you with your I-BiDaaS experience today?</p> <table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th></th> <th>Very dissatisfied</th> <th>Somewhat dissatisfied</th> <th>Neither satisfied nor dissatisfied</th> <th>Somewhat satisfied</th> <th>Very satisfied</th> </tr> </thead> <tbody> <tr> <td>Overall impression</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>User guidance and usability</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Information output (results)</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Throughput</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> </tbody> </table>		Very dissatisfied	Somewhat dissatisfied	Neither satisfied nor dissatisfied	Somewhat satisfied	Very satisfied	Overall impression	<input type="radio"/>	User guidance and usability	<input type="radio"/>	Information output (results)	<input type="radio"/>	Throughput	<input type="radio"/>																												
	Very dissatisfied	Somewhat dissatisfied	Neither satisfied nor dissatisfied	Somewhat satisfied	Very satisfied																																						
Overall impression	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
User guidance and usability	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Information output (results)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
Throughput	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>																																						
<p>10. How will you rate the I-BiDaaS application to the telecommunication sector?</p>																																											
<p>11. I-BiDaaS team would like to contact you again, if you agree, please provide the following information:</p> <p><input type="text" value="email address"/></p> <p><input type="text" value="Entrez votre réponse"/></p>																																											

Questionnaire for the CaixaBank Workshop: I-BiDaS application to the financial sector

<p>Section 1</p> <p>General information</p> <p>1. Please indicate your role in your organisation:</p> <p><input type="text" value="Sélectionnez votre réponse"/></p> <p>2. Do you work at CaixaBank S.A? If no, can you please indicate the name of your organisation?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> Autre <input type="text" value=""/></p> <p>3. Please indicate in which department do you work:</p> <p><input type="text" value="Entrez votre réponse"/></p> <p>4. Do you have any decision making power related to the investment on Big Data?</p> <p><input type="text" value="Sélectionnez votre réponse"/></p> <p>8. Which capabilities do you look for in a Big Data Analytics solution? Please select more than one if needed:</p> <p><input type="checkbox"/> Do it yourself as-a-service solution <input type="checkbox"/> Break data silos <input type="checkbox"/> Safe environment <input type="checkbox"/> Use wide range of analytical tools <input type="checkbox"/> Increase speed of data analysis <input type="checkbox"/> Cope with rate data asset growth <input type="checkbox"/> Autre <input type="text" value=""/></p> <p>9. Please indicate your level of agreement with the following statement:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>Strongly Disagree (1)</th> <th>Disagree (2)</th> <th>Neutral (3)</th> <th>Agree (4)</th> <th>Strongly Agree (5)</th> </tr> </thead> <tbody> <tr> <td>I am familiar with Big Data Analytics</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>I am familiar with Circular Economy</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Big Data Analytics can enable Circular Economy implementation</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>We are currently applying Circular Economy initiatives in my department</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>We plan to apply Circular Economy initiatives in the short-term</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> </tbody> </table> <p>10. Which Big Data Analytics tools do you use in your current work?</p> <p><input type="text" value="Entrez votre réponse"/></p>		Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)	I am familiar with Big Data Analytics	<input type="radio"/>	I am familiar with Circular Economy	<input type="radio"/>	Big Data Analytics can enable Circular Economy implementation	<input type="radio"/>	We are currently applying Circular Economy initiatives in my department	<input type="radio"/>	We plan to apply Circular Economy initiatives in the short-term	<input type="radio"/>	<p>5. Do you have any decision making power related to the selection of relevant technologies?</p> <p><input type="text" value="Sélectionnez votre réponse"/></p> <p>6. Do you know what is your organisations' level of investment in commercial products related to Big Data Analytics per year?</p> <p><input type="text" value="Sélectionnez votre réponse"/></p> <p>7. Which aspects regarding business performance does your organisation aim to optimise by using Big Data Analytics? Please select more than one if needed:</p> <p><input type="checkbox"/> Revenue Increase <input type="checkbox"/> Profit Increase <input type="checkbox"/> Cost Reduction <input type="checkbox"/> Time efficiency <input type="checkbox"/> Customer satisfaction <input type="checkbox"/> Product/Service Quality <input type="checkbox"/> Business Model Innovation <input type="checkbox"/> Launch new products/services <input type="checkbox"/> Data accessibility</p> <p>Section 2</p> <p>I-BiDaS experiments</p> <p>Use synthetic data</p> <p>11. Taking into account the presented use case using synthetic data, how feasible you see the usage of it for your potential use cases? (From 1 to 5):</p> <p>1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input type="radio"/> 5 <input type="radio"/></p> <p>12. Taking into account the previous question, please shortly describe the use case(s) you have in mind?</p> <p><input type="text" value="Entrez votre réponse"/></p> <p>13. Which aspects do you think synthetic data can have a greater impact?</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th>poor (1)</th> <th>Average (2)</th> <th>Good (3)</th> <th>Very good (4)</th> <th>Excellent (5)</th> </tr> </thead> <tbody> <tr> <td>Time efficiency</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Data accessibility</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>End-to-end execution time</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Accuracy and reliability of the analytical process</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> <tr> <td>Infrastructure cost</td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> <td><input type="radio"/></td> </tr> </tbody> </table> <p>14. Would you add any other functionality to the use case?</p> <p><input type="text" value="Entrez votre réponse"/></p>		poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)	Time efficiency	<input type="radio"/>	Data accessibility	<input type="radio"/>	End-to-end execution time	<input type="radio"/>	Accuracy and reliability of the analytical process	<input type="radio"/>	Infrastructure cost	<input type="radio"/>																																								
	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)																																																																				
I am familiar with Big Data Analytics	<input type="radio"/>																																																																								
I am familiar with Circular Economy	<input type="radio"/>																																																																								
Big Data Analytics can enable Circular Economy implementation	<input type="radio"/>																																																																								
We are currently applying Circular Economy initiatives in my department	<input type="radio"/>																																																																								
We plan to apply Circular Economy initiatives in the short-term	<input type="radio"/>																																																																								
	poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)																																																																				
Time efficiency	<input type="radio"/>																																																																								
Data accessibility	<input type="radio"/>																																																																								
End-to-end execution time	<input type="radio"/>																																																																								
Accuracy and reliability of the analytical process	<input type="radio"/>																																																																								
Infrastructure cost	<input type="radio"/>																																																																								

Section 3

I-BiDaaS experiments
Anomalies detection in the bank transfers from bank offices -financial terminal of the employees

15. Taking into account the presented use case "anomalies detection in the bank transfers from bank offices -financial terminal of the employees", how feasible you see the usage of it for your potential use cases? (From 1 to 5):

1	2	3	4	5
<input type="radio"/>				

16. Taking into account the previous question, please shortly describe the use case(s) you have in mind?

Entrez votre réponse

Section 4

I-BiDaaS experiments
Anomalies detection in mobile to mobile bank transfers

19. Taking into account the presented use case "Anomalies detection in mobile to mobile bank transfers", how feasible you see the usage of it for your potential use cases? (From 1 to 5):

1	2	3	4	5
<input type="radio"/>				

20. Taking into account the previous question, please shortly describe the use case(s) you have in mind?

Entrez votre réponse

21. Which aspects do you think I-BiDaaS solution can have a greater impact? (in the context of anomalies detection in mobile to mobile bank transfers):

	poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)
Time efficiency	<input type="radio"/>				
Data accessibility	<input type="radio"/>				
End-to-end execution time	<input type="radio"/>				
Accuracy and reliability of the analytical process	<input type="radio"/>				
Infrastructure cost	<input type="radio"/>				

17. Which aspects do you think I-BiDaaS solution can have a greater impact? (in the context of anomalies detection in the bank transfers from bank offices -financial terminal of the employees-).

	poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)
Time efficiency	<input type="radio"/>				
Data accessibility	<input type="radio"/>				
End-to-end execution time	<input type="radio"/>				
Accuracy and reliability of the analytical process	<input type="radio"/>				
Infrastructure cost	<input type="radio"/>				

18. Would you add any other functionality to the use case?

Entrez votre réponse

22. Would you add any other functionality to the use case?

Entrez votre réponse

23. Which aspects of the platform fulfil your needs? Please select more than one if needed:

- Batch analytics
- Stream analytics
- Advanced visualisation
- Support developers
- Support non IT-experts
- Autre

Section 5

Evaluation of end-to-end I-BiDaaS platform
Self-service mode

24. How easy and intuitive did you find using the I-BiDaaS platform when it comes to:

	Poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)
Project setup	<input type="radio"/>				
Selecting a data source	<input type="radio"/>				
Dataset fabrication	<input type="radio"/>				
Algorithm selection and setup	<input type="radio"/>				
Results Visualisation	<input type="radio"/>				

28. Are there any elements/features missing from I-BiDaaS? If yes, please specify:

Entrez votre réponse

29. How would you rate the potential of I-BiDaaS solution in the following aspects?

	Poor (1)	Average (2)	Good (3)	Very good (4)	Excellent (5)
Operability (real business setting possibilities)	<input type="radio"/>				
Innovation (bring new features with regards to other data analytics solutions)	<input type="radio"/>				
Compliance (assurance compliance with the current security and privacy regulations)	<input type="radio"/>				
Privacy Awareness (ensure the privacy of the data for getting out of company premises without business implications)	<input type="radio"/>				
Cost reduction (To what extent I-BiDaaS solution could be cost-effective compared to current solutions)	<input type="radio"/>				

27. Innovation in the business-related sense is defined as something that provides a specific value that no other product or service could offer before. Do you think I-BiDaaS has such elements or features? If yes, please name and describe them:

Entrez votre réponse

30. How will you rate the I-BiDaaS application to the financial sector?

☆ ☆ ☆ ☆ ☆