



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

### **D6.5: Assessment Report and Impact Analysis<sup>†</sup>**

**Abstract:** This deliverable provides a critical assessment and impact analysis of the final I-BiDaaS solution. Considering the evaluation methodology defined in Deliverable 1.3 [1], all KPIs are reported for each component of the platform both in operational terms (cost, service levels, etc.) and technical terms (performance of solution). This report takes into consideration the implementation of the real-life industrial experiments of task 6.2 and the evaluation and impact analysis of task 6.3. Furthermore, the deliverable reports on the impact analysis and external stakeholders' involvement during the progress of the project from M33 to M36, and also documents directions for exploitation and potential commercialization.

Contractual Date of Delivery	31/12/2020
Actual Date of Delivery	31/12/2020
Deliverable Security Class	Public
Editor	<i>Ioannis Arapakis (TID)</i>
Contributors	All I-BiDaaS partners
Quality Assurance	<i>Giuseppe Danilo Spennacchio (CRF)</i> <i>Leonidas Kallipolitis, Efstathios Dimakos (AEGIS)</i> <i>Kostas Lampropoulos (FORTH)</i>

---

<sup>†</sup> 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. *Giuseppe Danilo Spennacchio (CRF)*
2. *Leonidas Kallipolitis, Efstathios Dimakos (AEGIS)*
3. *Kostas Lampropoulos (FORTH)*

### Revisions

Version	Date	By	Overview
0.1	06/11/2020	TID	Table of Contents
0.2	17/11/2020	TID	Consolidate Table of Contents
0.3	06/12/2020	TID	1 <sup>st</sup> integrated document with contributions from FORTH, BSC, IBM, SAG, UNIMAN, ITML, UNSPMF, CAIXA, CRF and AEGIS
0.4	20/12/2020	TID	2 <sup>nd</sup> integrated document with contributions from FORTH, BSC, IBM, SAG, UNIMAN, ITML, UNSPMF, ENPC, CAIXA, CRF and AEGIS
0.5	28/12/2020	TID	3 <sup>rd</sup> integrated document with contributions from FORTH, ATOS, ITML, UNSPMF, UNIMAN and CRF
0.6	30/12/2020	TID	Integrated version after 1 <sup>st</sup> review of IRs (CRF and AEGIS)
0.7	31/12/2020	TID	Final Version

## Table of Contents

<b>LIST OF FIGURES .....</b>	<b>5</b>
<b>LIST OF TABLES .....</b>	<b>7</b>
<b>LIST OF ABBREVIATIONS .....</b>	<b>8</b>
<b>EXECUTIVE SUMMARY .....</b>	<b>9</b>
<b>1. INTRODUCTION .....</b>	<b>10</b>
<b>2. FINAL I-BIDAAS SOLUTION.....</b>	<b>11</b>
<b>3. I-BIDAAS PLATFORM EVALUATION .....</b>	<b>13</b>
3.1. INDIVIDUAL PARTS EVALUATION & BENCHMARKING .....	13
<i>Test Data Fabrication (TDF)</i> .....	13
<i>Universal Messaging</i> .....	14
<i>Advanced ML</i> .....	14
<i>CEP Engine</i> .....	15
<i>GPU-accelerated pattern matching</i> .....	15
<i>Hecuba and QBeast</i> .....	16
<i>Resource Management and Orchestration module</i> .....	17
<i>Visualisation Tool</i> .....	18
3.2. OVERALL I-BIDAAS PROTOTYPE QUALITY TESTS.....	20
<i>Testing Methodologies and Procedure</i> .....	21
3.3. USE CASES EVALUATION.....	22
<i>Telecommunications experiments</i> .....	22
<i>Banking experiments</i> .....	25
<i>Manufacturing experiments</i> .....	29
<i>Generic experiments</i> .....	32
3.4. DEVIATION FROM EXPECTED RESULTS .....	35
<b>4. COMPLIANCE WITH GENERIC BIG DATA PIPELINES AND BLUEPRINTS .....</b>	<b>39</b>
<b>5. IMPACT ANALYSIS .....</b>	<b>41</b>
5.1. INDUSTRIAL IMPACT ASSESSMENT .....	41
5.2. FEEDBACK FROM EXTERNAL STAKEHOLDERS.....	60
5.3. EXPLOITATION AND POTENTIAL COMMERCIALIZATION .....	67
<b>6. CONCLUSIONS.....</b>	<b>69</b>
<b>7. REFERENCES .....</b>	<b>70</b>
<b>ANNEX I – WALK-THROUGH EXPERT &amp; SELF-SERVICE MODE .....</b>	<b>71</b>
<b>ANNEX II – STANDARDISATION ACTIVITIES .....</b>	<b>91</b>

## List of Figures

Figure 1. Information on experiment duration.....	19
Figure 2. Visualisation of Control to Online Banking .....	20
Figure 3. Throughput and latency measurements .....	20
Figure 4. Sample of the ‘Enhanced Control of customers to Online Banking’ use case clustering results in the I-BiDaaS platform .....	26
Figure 5. (a) DBSCAN representation, (b) K-means representation.....	27
Figure 6. Heat-plot of the 84 most relevant attributes from the 501 original attributes.....	28
Figure 7. Results for 100 random sampling taken from the real and the synthetic data (5K datapoints each) and the pMSE calculated using a logistic model .....	29
Figure 8. Sensor category selection and visualisation of anomalies .....	30
Figure 9. Live chart with the trend of parameters for the last 20 engine blocks.....	31
Figure 10. Colour-coded table, pie-chart and parameters statistics.....	32
Figure 11. The actual difference of Sentiment and CSI scores for each of the transcripts in our dataset.....	36
Figure 12. The accumulated number of transcripts for a given percentage difference. ....	36
Figure 13. Confusion matrix of the overall performance. ....	37
Figure 14. Mapping of I-BiDaaS architecture to the DataBench pipeline.....	39
Figure 15. Alignment between I-BiDaaS banking experimental workflow and the DataBench pipeline in the context of the banking experiments.....	40
Figure 16. Screenshot of AWS servers used for the purposes of the Telefonica Research Hackathon. 61	
Figure 17. Feedback to the question "Rate your event experience (being 5 the best rate experience)" 62	
Figure 18. EBDVF2020 - Geographical Spread .....	63
Figure 19. To which of our stakeholder types do you belong? .....	64
Figure 20. What is the main barrier or risk preventing you from implementing Big Data Analytical solutions in your organisation? .....	64
Figure 21. I-BiDaaS Final Event - Geographical Spread .....	66
Figure 22. To which of our stakeholder types do you belong?.....	66
Figure 23. What is the main barrier or risk preventing you from implementing Big Data Analytical solutions in your organisation? .....	66
Figure 24. Main page .....	71
Figure 25. Expert mode.....	72
Figure 26. Preparing the project (I).....	72
Figure 27. Code uploading .....	73
Figure 28. Experiment details.....	73
Figure 29. Running experiment .....	74
Figure 30. Project details page.....	75
Figure 31. Project details (II).....	75
Figure 32. Experiment Results page .....	76
Figure 33. Output.txt contents .....	77
Figure 34. Sample of the successful execution of random forest template .....	77
Figure 35. I-BiDaaS main page .....	79
Figure 36. Self-Service Mode page.....	79
Figure 37. I-BiDaaS questionnaire for selecting the appropriate algorithm.....	80
Figure 38. K-means Prediction – at a glance .....	80
Figure 39. K-means Prediction – Project Details page.....	81
Figure 40. K-means Prediction – Experiment details.....	81
Figure 41. K-means Prediction – data preview.....	82
Figure 42. K-means Prediction – select Computational Resources .....	82
Figure 43. K-means Prediction – Running experiment .....	82
Figure 44. K-means Prediction – Project details page .....	83
Figure 45. K-means Prediction – Project details.....	83
Figure 46. K-means Prediction – Visualise results .....	84
Figure 47. K-means Evaluation – Project Details page.....	85
Figure 48. K-means Evaluation – Experiment details.....	85
Figure 49. K-means Evaluation – data preview.....	86

Figure 50. K-means Evaluation – Project details page.....	86
Figure 51. K-means Evaluation – Project details page.....	87
Figure 52. K-means Evaluation – Visualise results .....	88
Figure 53. ADMM Lasso – Quick preview .....	88
Figure 54. ADMM in I-BiDaaS – How it works and a usage example – .....	89
Figure 55. ADMM Lasso – Project details page.....	89
Figure 56. ADMM Lasso – Visualise results.....	90

## List of Tables

Table 1: Summary of the I-BiDaaS use cases available for the Co-Develop mode.....	11
Table 2: Test Data Fabrication component.....	13
Table 3: Universal Messaging component .....	14
Table 4: Advanced Machine Learning sub-module .....	14
Table 5: Complex Event Processing component .....	15
Table 6: Graphics Processing Unit component.....	16
Table 7: Hecuba DBS component.....	16
Table 8: Qbeast component .....	16
Table 9: Resource Management and Orchestration Module component.....	17
Table 10: Visualisation Tool component.....	18
Table 11: Overall I-BiDaaS prototype evaluation results.....	21
Table 12: Quality of Service in Call Centers.....	23
Table 13: Accurate Location Prediction with High Traffic and Visibility .....	23
Table 14: Performance of binary classification algorithms for the antenna KPIs use case .....	24
Table 15: Optimization of Placement of Telecommunication Equipment .....	25
Table 16: Enhance control of customers to online banking.....	25
Table 17: Advanced Analysis of bank transfer payment in financial terminal.....	27
Table 18: Analysis of relationships through IP address .....	28
Table 19: Maintenance and monitoring of production assets .....	29
Table 20: Production process of aluminium casting .....	30
Table 21: Questionnaire for the evaluation of the end-to-end I-BiDaaS platform in Self-Service mode .....	32
Table 22: Questionnaire for the evaluation of the end-to-end I-BiDaaS platform in Expert mode .....	33
Table 23: Experiment for end-to-end I-BiDaaS platform in Self-Service mode .....	34
Table 24: Experiment for end-to-end I-BiDaaS platform in Expert mode .....	35
Table 25: I-BiDaaS KPIs.....	41
Table 26: Explanation of the I-BiDaaS KPIs.....	42
Table 27: I-BiDaaS Tools & Technologies – Achieved Technology Readiness Level.....	50
Table 28: List of I-BiDaaS Excellent Innovations as accepted from EU Innovation Radar.....	52
Table 29: I-BiDaaS Use Cases – Achieved Technology Readiness Level.....	53
Table 30: Summary of the impact of the CAIXA use cases studied in I-BiDaaS.....	55
Table 31: Summary of the impact of the CRF use cases studied in I-BiDaaS.....	55
Table 32: Summary of the impact of the TID use cases studied in I-BiDaaS .....	56
Table 33: EAB Comments and how I-BiDaaS works towards addressing them .....	57
Table 34: List of Standards used in I-BiDaaS Technologies.....	91
Table 35: List of Standards used in the use-case experiments.....	91
Table 36: List of Standards per layer related to I-BiDaaS.....	92
Table 37: Analysis of standard processes and methodologies per use-case .....	93
Table 38: Participation to Standardisation Bodies per partner .....	94
Table 39: Mapping between requirements, activities and functional components .....	94

## List of Abbreviations

ADDM	Automatic Database Diagnostic Monitor
AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
ATM	Automated Teller Machine
BDV	Big Data Value
BDVA	Big Data Value Association
BI	Business Intelligence
CEP	Complex Event Processing
CRM	Customer Relationship Management
CSI	Customer Satisfaction Index
CSP	Constraint Satisfaction Problem
CSV	Comma-separated Values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
E2E	End-to-End
EAB	External Advisory Board
GPU	Graphics Processing Unit
I-BiDaaS	Industrial-Driven Big Data as a Self-Service Solution
IOP	In Out Processor
IoT	Internet of Things
IP	Internet Protocol
IT	Information Technology
JSON	JavaScript Object Notation
KNN	K-Nearest Neighbor
KPI	Key Performance Indicators
KVM	Kernel-based Virtual Machine
LASSO	Least Absolute Shrinkage and Selection Operator
LLC	Limited Liability Company
ML	Machine Learning
MPI	Message Passing Interface
MQTT	Message Queuing Telemetry Transport
MVP	Minimum Viable Product
NIST	National Institute of Standards and Technology
NoSQL	Not Structured Query Language
OLTP	Online Transaction Processing
PoC	Proof-of-Concept
PPP	Public-Private-Partnership
RAM	Random Access Memory
RMO	Resource Management and Orchestration
SME	Small and Medium-sized Enterprise
TCP	Transmission Control Protocol
TDF	Test Data Fabrication
UM	Universal Messaging
VM	Virtual Machine



## Executive Summary

This deliverable provides a concluding assessment of the final I-BiDaaS solution through a review of a series of real-life industrial experiments from the telecommunication, banking and manufacturing industries, which demonstrate the successful application of the I-BiDaaS solution in real-world environments. The I-BiDaaS project, funded by the Horizon 2020 Programme under Grant Agreement 780787, aims to empower both expert and non-expert users with an intuitive medium for interacting with Big Data technologies by designing, implementing, 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) intuitive, intelligent and effective visualization and interaction capabilities for the end-users.

All activities were guided by the I-BiDaaS experimental protocol to ensure the successful implementation of the operational experiments and their alignment with the key business objectives, as determined by the industrial partners for the reported use cases. Further revisions were taken into account regarding the real-life industrial experiments during the implementation phase and the design process of the I-BiDaaS platform and other associated technologies.

In more detail, the deliverable reports a comprehensive evaluation of the final I-BiDaaS platform and all of its main components through a series of exhaustive quantitative and qualitative industry-validated benchmarks that consider both operational (cost, service levels, etc.) and technical terms (performance of solution). The integration was performed by ensuring secure data management through the anonymization and encryption of the data, while several programming languages and advanced visualization tools have been used to develop a platform easy-to-use for all experiments, as demonstrated by a rigorous use case evaluation, in the context of the three addressed industrial sectors. Furthermore, the assessment of the I-BiDaaS platform is situated beyond the aforementioned use cases, and it is carried out in a wider network of industry peers and resources, such as the one offered by the European project DataBench, as well as with the support of external stakeholders. Last, the deliverable provides a critical discussion on the achieved impact (by concluding the work started in D6.2 [2], D6.3 [3] and D6.4 [4]) with respect to the envisioned project innovation, and charts future directions for exploitation and potential commercialization on a global scale.

# 1. Introduction

This deliverable provides a final, comprehensive assessment of the I-BiDaaS solution, and all of its main components, and brings to a conclusion the work done in “WP6. Real – life industrial and operational experiments”. Following the experimental protocol alignment (Task 6.1), it reports on a series of real-life industrial experiments from the telecommunication, banking and manufacturing industries (Task 6.2), and demonstrates 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 for both batch- and real-time processing modes. Batch and stream-processing, discussed in more 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. These aspects include cases that require a deeper analysis of large volumes of data, collected over a period of time (batch), or those that require velocity and agility for the events that we are called 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 activities were guided by the I-BiDaaS experimental protocol to ensure the successful implementation of the operational experiments and their alignment with the key business objectives and KPIs for each experiment. The I-BiDaaS solution aims to demonstrate in a realistic, measurable, and replicable way the effects that Big Data have on different real scenarios and also empower both expert and non-expert users with an intuitive medium for interacting with big data technologies offering 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. 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.

Furthermore, this deliverable reports on the progress of the impact analysis with respect to the expected project level innovation and achievements and provides a description of the activities that involved external stakeholders who have expertise, experience or interest in Big Data analytics in the evaluation process. Last, the deliverable provides a critical discussion on the achieved with respect to the envisioned project innovation, and charts future directions for exploitation and potential commercialization on a global scale.

The rest of the document is structured as follows. Section 2 provides a summary of the final I-BiDaaS solution. Section 3 provides a detailed evaluation of the final I-BiDaaS solution and its main components, through a series of exhaustive quantitative and qualitative industry-validated benchmarks that consider both operational and technical terms, as well as a detailed discussion of the use cases evaluation. Section 4 discusses the project’s compliance with generic big data pipelines and blueprints. Section 5 provides an assessment over the achieved industrial impact, incorporates the feedback of external stakeholders, and also discusses relevant exploitation and commercialization activities. Section 6 concludes with a summary of the results of WP6 achieved from M19 to M36.

## 2. Final I-BiDaaS solution

The functionality of the I-BiDaaS solution, described in Section 2 and Section 3 of D5.6 [5], can be divided into three categories with the following characteristics:

- The **Expert mode** allows experts (e.g., python developers with ML expertise) to upload their own data analytics code based on the available and highly reusable I-BiDaaS templates.
- The **Self-service mode** allows industry non-expert, in-house personnel, with domain knowledge and partial insights about data analysis, to run effortlessly Big Data analytics tasks in a user-friendly manner, selecting a pre-defined ML algorithm from a list of options.
- The **Co-Develop mode** corresponds to an end-to-end solution for a given industry project developed by the I-BiDaaS team (i.e. the I-BiDaaS use cases).

In the Self-service and the Expert mode, users can either initiate existing projects (Self-service mode) with already implemented algorithms, e.g., K-means, or create custom ones by uploading and running their own code (Expert mode). Then, they create and execute their experiment(s) based on their data and preferred configuration. These processes are discussed in depth in D5.6 [5], section 2.1. With respect to the Co-Develop mode, the I-BiDaaS Consortium implemented 8 use cases, summarized in Table 1. More information about these use cases can be found in D5.6 [5], chapter 3, and D6.4 [4].

**Table 1: Summary of the I-BiDaaS use cases available for the Co-Develop mode**

Use case name	Description	Type
<b>Accurate location prediction with high traffic and visibility</b>	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.	Batch processing
<b>Optimization of Placement of Telecommunication Equipment</b>	This use case addresses the problem of improving the routing and placement of the telecommunication equipment that is already in place or arrange accordingly the new equipment obtained, based on an automatic analysis of usage data retrieved by the customers.	Stream processing
<b>Quality of Service in Call Centers</b>	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.	Stream processing
<b>Enhanced control of customers to online banking</b>	In this use case, we focused on analysing 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.	Batch & Stream processing
<b>Advanced analysis of bank transfer payment in financial terminal</b>	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).	Batch processing

<b>Analysis of relationships through IP addresses</b>	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.	Batch & Stream processing
<b>Maintenance and monitoring of production assets</b>	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 failures before they happen by doing maintenance at the right time (not too late or too early, to avoid inefficiencies).	Batch processing
<b>Production process of aluminium die-casting</b>	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. By monitoring in real-time parameters, it is possible to check their trend and to visualise the classification levels of the quality of the engine block to avoid further processing and scraps, thanks to the possibility to act before the next production steps.	Stream processing

### 3. I-BiDaaS Platform Evaluation

This section discusses the outcome of the I-BiDaaS platform evaluation activities conducted in WP6. To this end, we review each major component in isolation and report the relevant experimental indicators and associated metrics.

#### 3.1. Individual parts evaluation & benchmarking

##### Test Data Fabrication (TDF)

The TDF component (Table 2) was installed on a dedicated Virtual Machine (VM). The latest available version of the tool (1.0.2.3) is installed and was used for the evaluations. The TDF 1.0.2.3 version includes the following new features and support: New Modeler UI, TDF parallel edition (parallel PRB solver), Support for docker container, Support for MariaDB, Performance bug fixes, MS SQL check constraints support, Unsupported data types checks, Single foreign key to multiple primary keys support, Extended REST API. Additional software modules that communicate with TDF (e.g., a parallel CSP solver, an SQLite database and a PostgreSQL database) were also installed and updated. In addition to the TDF server, the installation included the user management module, the resource management module, and a license service module to complement all the requirements for the TDF service. The TDF tests are discussed in detail in deliverables D6.4 [4] and D6.1 [6].

**Table 2: Test Data Fabrication component**

TDF (IBM)			
Indicator	Metric	Benchmark (if applicable) <sup>1</sup>	Measurement obtained
<b>Scalability</b>	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
<b>Validity</b>	Generated data must fit the data model	The data fits the model	Data fits the model
<b>Performance</b>	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
<b>Accuracy</b>	Measured against real data	Evaluate the quality of the fabricated data by comparing analyses results of both real and fabricated data. and also by evaluating the pMSE score of the synthetic data	Data was manually examined by relevant partners, compared by applying data analysis on both datasets and verifying distributions and parameters similarities (specific utility). The mean pMSE score for the IP dataset is 0.234 with a STD of 0.000835, and the measured pMSE score for the ‘Aluminium die-casting’ dataset is 0.218 with a STD of 0.00146 (general utility).
<b>Availability</b>	No of crashes	No crashes	No crashes observed during the runs performed

<sup>1</sup> For the non-applicable: Units are included instead of benchmarks.

## Universal Messaging

The Universal Messaging (UM) component is an Enterprise Message Broker that provides high throughput of messages and supports many industry-grade standards like JMS, AMQP and MQTT. In the final I-BiDaaS solution, we use MQTT to deliver and distribute messages to the different components.

The numbers shown in Table 3 were measured on a local, stand-alone high-performance laptop. However, one can expect that the values are even higher when using a dedicated machine in a cloud or on-premise data center.

**Table 3: Universal Messaging component**

Universal Messaging (SAG)			
Indicator	Metric	Benchmark (if applicable) <sup>2</sup>	Measurement obtained
<b>Scalability</b>	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
<b>Operational performance</b>	Execution time	seconds	ms
	Latency	seconds	ms
<b>Reliability</b>	Data failure	N/A	no failure
<b>Compliance</b>	Measured against relevant standards	JMS, MQTT, AMQP	in line with standards

## Advanced ML

The performance of the Advanced Machine Learning component has been carried out according to the methodology defined in D6.2 [2]. Table 4 below shows a summary of the achieved results, as well as pointers to other I-BiDaaS deliverables where the details of the respective measurements can be found.

**Table 4: Advanced Machine Learning sub-module**

Advanced ML (UNSPMF)			
Indicator	Metric	Benchmark (if applicable) <sup>3</sup>	Measurement obtained
<b>Quality of results</b>	For optimization solver: Objective function value; constraints violation value	Standard solver on the same problem, if feasible for the given dataset and if implementation is available: CVXPY <sup>4</sup> , 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 [3].
	For supervised learning: Training error; testing error; classification accuracy	Benchmark: Sklearn <sup>5</sup> , for moderate size data. Target: have a comparable result with the benchmark, with improved scalability.	See Tables 5-7 in D3.1 [7]; Table 7 in D6.3 [3]; and Subsection 4.1.1.3 and Figure 13 in D3.3 [8], and the associated descriptions.

<sup>2</sup> For the non-applicable: Units are included instead of benchmarks.

<sup>3</sup> For the non-applicable: Units are included instead of benchmarks.

<sup>4</sup> Python-embedded modelling language for Convex optimization problems, available at <https://www.cvxpy.org/>

<sup>5</sup> <https://scikit-learn.org/stable/index.html>

	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 [3] and the associated descriptions.
<b>Scalability</b>	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 [3], and Figure 13 in D3.3 [8], and the associated descriptions.
<b>Performance</b>	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 [8] and Subsection 4.1.1.3, second paragraph of D3.3 [8]

## CEP Engine

In order to support real-time analytics in the I-BiDaaS solution, we use the Apama streaming analytics platform. It can directly read data from message brokers like UM or from MQTT brokers or Kafka. Apama applies complex event processing to the incoming data streams to look for anomalies and initiate pre-defined actions.

The performance analysis shown in Table 5 was held on a local, stand-alone high-performance laptop machine. However, one should expect that the performance metrics will be even higher if the benchmarking is carried out on a dedicated machine in the cloud or in-premise data center.

**Table 5: Complex Event Processing component**

CEP Engine (SAG)			
Indicator	Metric	Benchmark (if applicable) <sup>6</sup>	Measurement obtained
<b>Scalability</b>	Response time	seconds	ms
	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

The GPU-accelerated pattern matching is discussed in detail in deliverables D4.1 [11] and D4.2 [12]. In essence, the GPU-accelerated pattern matching operations are offered through an OpenCL library that supports both string searching and regular expression matching operations. The library provides a C/C++ API for processing incoming records, and returning any matches found back to the application. As described in D4.3 [13], these operations are used to provide

<sup>6</sup> For the non-applicable: Units are included instead of benchmarks.

different types of text analytics (such as sentiment score, word frequencies, and most frequent words).

**Table 6: Graphics Processing Unit component**

GPU-accelerated pattern matching (FORTH)			
Indicator	Metric	Benchmark (if applicable) <sup>7</sup>	Measurement obtained
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 [3]).
	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 [3]).

## Hecuba and QBeast

The Hecuba and Qbeast tests are discussed in detail in deliverable D6.4 [4]. In Table 7 and Table 8, we provide a summary of those.

**Table 7: Hecuba DBS component**

Hecuba DBS (BSC)			
Indicator	Metric	Benchmark (if applicable) <sup>8</sup>	Measurement obtained
Scalability	Speedup	Benchmark: Dislib <sup>9</sup> Baseline: Dislib using files	2.54 max (for K-means)
Operational Performance	Response time		68 % from the original (in KNN)
	IOPS		N/A
	Disk usage		N/A
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)

**Table 8: Qbeast component**

Qbeast component Qbeast (BSC)			
Indicator	Metric	Benchmark (if applicable) <sup>10</sup>	Measurement obtained
Effect on Machine learning algorithms	Speedup	Benchmark: Dislib and MLlib <sup>11</sup> 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

<sup>7</sup> For the non-applicable: Units are included instead of benchmarks.

<sup>8</sup> For the non-applicable: Units are included instead of benchmarks.

<sup>9</sup> The Distributed Computing Library, available at <https://dislib.bsc.es/>

<sup>10</sup> For the non-applicable: Units are included instead of benchmarks.

<sup>11</sup> The Apache Spark's scalable machine learning library, available at <https://spark.apache.org/mllib/>



			“All 0.01%”, we have the highest speedup as we benefit the most from the efficient sampling of Qbeast (see Table 22 in D6.4 [4]).
<b>Scalability</b>		Benchmark: synthetic geographical queries  Baseline: PostGis <sup>12</sup>	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 in D6.4 [4])
<b>Operational performance</b>	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 Figures 40-41 in D6.4 [4])
	IOPS		With two nodes, Qbeast, achieves 83K IOPS (performing as Cassandra), and approximately improves 80% when doubling the nodes (see Figure 39 in D6.4 [4])
	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.
<b>Availability</b>	% timeouts		0%
<b>Reliability</b>	Fault tolerance on down nodes (amount of missed data and slowdown during the recovery process)		Fault tolerance mechanisms provided by Cassandra (e.g., replication)

## Resource Management and Orchestration module

The evaluation results of the RMO module have been reported in D6.4 [4], see table below. The RMO has been tested against private and public Cloud Service Providers, as well as over Edge environments (Raspberry phi, Jetson nano). Different deployment topologies have been evaluated including VMs, docker containers and K8s clusters.

**Table 9: Resource Management and Orchestration Module component**

Resource management and orchestration module (ATOS)			
Indicator	Metric	Benchmark (if applicable) <sup>13</sup>	Measurement obtained
<b>Operational performance</b>	Response time	- Page response time (in seconds) - Transactions processed	100-150 ms

<sup>12</sup> A spatial database extender for PostgreSQL object-relational database, available at <https://postgis.net/>

<sup>13</sup> For the non-applicable: Units are included instead of benchmarks.

		- Network error, latency & utilization	
	Resource utilization	Delay between a client request and a cloud service provider's response. (in seconds)	< 1 second
<b>Scalability</b>	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 flavour (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.
<b>Availability</b>	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.
<b>Reliability</b>	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

The evaluation results of the Visualisation Tool have been reported analytically in D6.4 [4], where the performance of the tool has been evaluated during the final experiment implementation. Table 10 below provides a summary of this report and shows the obtained quantitative evaluation results.

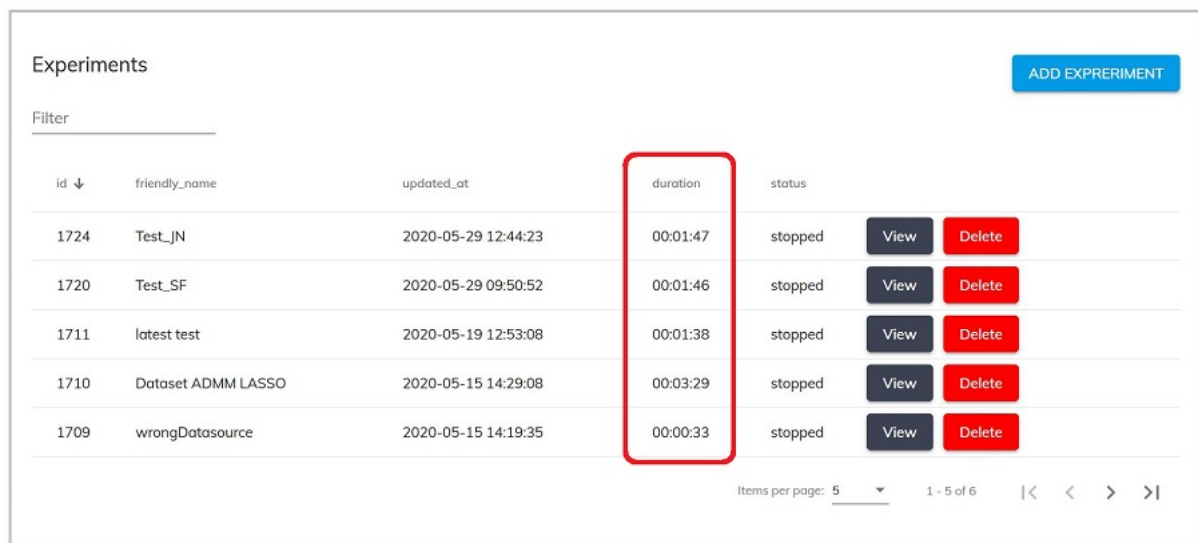
**Table 10: Visualisation Tool component**

Visualisation Tool (AEGIS)			
Indicator	Metric	Benchmark (if applicable) <sup>14</sup>	Measurement obtained
<b>Operational performance</b>	Response time	7-10s	Initial Page load: 9.92s (avg) Internal Page Loading: <1.25s (avg)
<b>Availability</b>	Uptime	95%	No downtime was experienced during the operation of the platform. Platform was only not available during planned maintenance and version updating.
<b>Reliability</b>	Fault Tolerance	Interface responsive in case of data errors.	No errors were detected that prevented interface from being responsive. Faulty

<sup>14</sup> For the non-applicable: Units are included instead of benchmarks.

		Informative messages to users.	information is not permitted by the interface.
<b>Usability (Efficiency)</b>	Perception of task completion quality	>80% positive perception. Results via 1-5 scaled questions.	87.5% achieved
<b>Usability (Satisfaction)</b>	Degree to which user needs are satisfied - look and feel	>80% positive perception. Results via 1-5 scaled questions.	87.5% achieved

Work performed after that period and up until the end of the project includes some further fine-tuning of visual elements available in the various use cases. Towards these improvements, the user-centered design approach was followed (as in all the design process) so as to add enhancements that would increase useful information displayed, while keeping the UI consistent with existing UI elements. Such an example is the addition of information on an experiment's duration, as shown in the overview list of experiments inside a project (Figure 1).

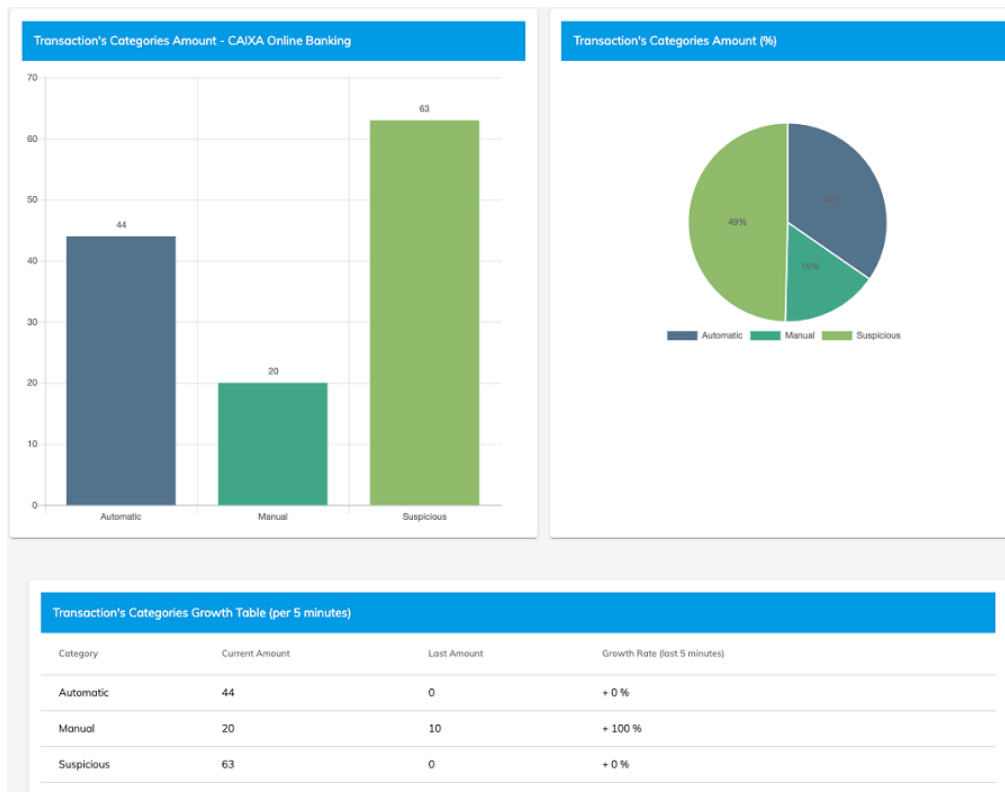


Experiments					ADD EXPERIMENT	
Filter						
id ↓	friendly_name	updated_at	duration	status		
1724	Test_JN	2020-05-29 12:44:23	00:01:47	stopped	View	Delete
1720	Test_SF	2020-05-29 09:50:52	00:01:46	stopped	View	Delete
1711	latest test	2020-05-19 12:53:08	00:01:38	stopped	View	Delete
1710	Dataset ADMM LASSO	2020-05-15 14:29:08	00:03:29	stopped	View	Delete
1709	wrongDatasource	2020-05-15 14:19:35	00:00:33	stopped	View	Delete

Items per page: 5 1 - 5 of 6 |< < > >|

**Figure 1. Information on experiment duration**

Furthermore, taking advantage of the expandability of the tool and the seamless integration with the I-BiDaaS platform, we managed to additionally develop the visualization for the use case “Enhanced control of customers to online banking”. The data for this use case comes from a stream, so the experience from other use cases providing streaming data was valuable to quickly setup the visualization and integrate it as a complete example in the “Co-Develop” mode of the platform. Figure 2 below shows a snapshot of the visualization where all the widgets are updated in real-time as the data stream is received, thus allowing users to have a dynamic overview of the underlying analysis through an intuitive and dynamic visualization.

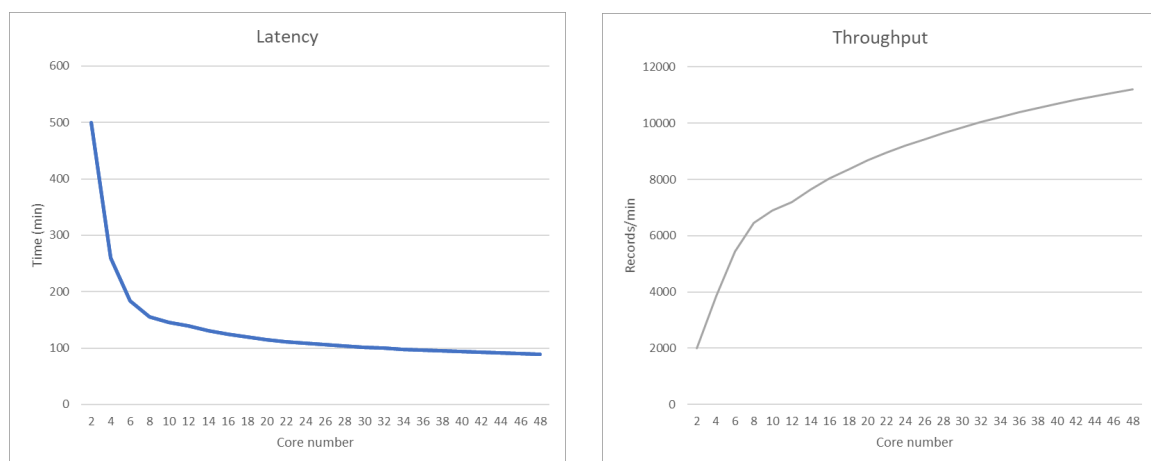


**Figure 2. Visualisation of Control to Online Banking**

### 3.2. Overall I-BiDaaS prototype quality tests

For benchmarking the I-BiDaaS platform, we chose to measure the throughput and latency using two of the available Machine Learning (ML) in the Self-Service mode: K-means and Random Forest. These are two of the most common examples of unsupervised and supervised learning. Both algorithms have been implemented in PyCOMPSs, using the Dislib library.

In our experiments (Figure 3), latency is calculated as the run time of the experiment (in minutes), and throughput is calculated as number of records per minute. We run our experiments in a docker swarm comprised of 48 cores, providing 2GB of memory per worker/core. To test the two algorithms, a synthetic dataset of 1.000.000 records was created. For the Random Forest, each record of the dataset consists of two scalar fields and for the K-means the dataset consists of two scalar fields and a label.



**Figure 3. Throughput and latency measurements**

More large-scale experiments of PyCOMPSs, in comparison with other Big data batch processing technologies and also including scalability metrics such as speedup, can be found in related research [9].

Moreover, experiments validated that the penalty of initializing and running the PyCOMPSs code through the I-BiDaaS AVT is insignificant. Starting an experiment through the UI takes milliseconds, and then the same core executes as in the case of running PyCOMPSs standalone in a cluster. Table 11 provides updated information in relation to the corresponding table in D6.4 [4], with respect to latency, throughput and speedup.

**Table 11: Overall I-BiDaaS prototype evaluation results**

Indicator	Metric	Benchmark	Measurement obtained
<b>Usability</b>	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
<b>Scalability</b>	Speedup	Benchmark: Dislib <sup>15</sup> Baseline: Dislib using files	Identical to dislib
<b>Operational Performance</b>	Response time (Latency)		Identical to dislib
	Data throughput (IOPS, no of generated data records per time unit)		Identical to dislib
	Resources utilization (storage, memory, CPU)		Identical to dislib
<b>Availability</b>	Uptime, % timeout		Not measured
<b>Reliability</b>	Data failure, Fault tolerance		Not measured
<b>Data Security</b>	Compliance with relevant security and privacy regulations and standards	At least 90% compliance	100%
<b>Privacy</b>	Compliance with relevant security and privacy regulations and standards	At least 90% compliance	100%
<b>Compliance</b>	Measured against relevant standards	NA	4.5/5
<b>Cost</b>	Compared against commercial alternatives	>30% reduced costs with respect to competitive commercial solutions	4/5

## Testing Methodologies and Procedure

For the evaluation of the I-BiDaaS solution, several different tests were conducted.

- *Unit Testing*: For each component developed within I-BiDaaS, unit testing was performed to verify the functions created to implement the business logic for each component and also the results. (See Section 3.1)
- *UM Testing*: This type of testing handles communication between the major components of the I-BiDaaS platform for streaming the use cases. The purpose is to

<sup>15</sup> The Distributed Computing Library, available at <https://dislib.bsc.es/>

verify that all messages will be delivered, that each component will be able to receive messages and that each component will be able to send messages.

- *End-to-End Testing*: ITML successfully completed this type of testing (both in software and hardware terms) and verified the desired functionality across the whole platform, offering insights on the performance and in this case of a multi-component system demonstrating the intended communication between these components.

A more elaborate description of the tests and associated results is reported in deliverable D5.6 [5].

### 3.3. Use cases evaluation

This section provides a more exhaustive description of the results obtained in the experimentation with the different use cases.

#### Telecommunications experiments

For the “Quality of Service in Call Centers” use case, the final I-BiDaaS solution offers an integrated full-stack service for processing and extracting actionable knowledge from big data, that includes data capture and integration from a variety of different sources and formats, scalable batch and real-time data processing analytics. This solution involves the extraction of the transcripts from the audio calls, the generation of prosodic and linguistic features, and finally forwarding this meta-data as input to a predictive model that classifies customer satisfaction.

To this end, we implemented a series of tools that accept as input a processed audio file, and the corresponding transcripts, and output a Customer Satisfaction Index (CSI). More specifically, we analysed the significance of various acoustic features extracted from customer-agents’ spoken interaction in predicting self-reported satisfaction and investigated whether speech prosodic features can deliver complementary information to speech transcriptions provided by Automatic Speech Recognition (ASR). We then investigated the effectiveness of deep neural architectures (Convolutional Neural Networks) in performing early feature fusion on both prosodic and linguistic information, for the binary classification task of “low” and “high” satisfaction prediction. Our initial results, reported also in D6.4 [4], revealed that linguistic features can predict self-reported satisfaction more accurately than those based on prosodic and conversational descriptors. Also, the fusion of linguistic and prosodic features achieved the best performance in our experiments, suggesting the complementarity of the information conveyed by each set of behavioural representation. This pipeline was tested on real data, as well as synthetic data that have similar properties as the real data and can support the same quality of analysis, thus allowing us to deploy and benchmark the CC demo rapidly and timely, and without having to address the potential security and privacy risks.

Business Units inspect <1% of total calls per year. The current solutions are both costly and time-consuming. However, through a number of simple, intuitive, and effective visualizations and dashboards for the end-users, the I-BiDaaS platform can facilitate a more effective operation of the call centres (improvement of QoS, automation, etc.) and support mining capabilities for monitoring customer satisfaction. As shown in Table 12, the proposed service has the capacity to increase the number of low customer satisfaction audio calls detected by human agents to 7,000 (a notable 200% increase) by pre-processing/filtering the audio calls and flagging those cases of interest.

**Table 12: Quality of Service in Call Centers**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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).
<b>Application Level Performance indicators</b>	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).
<b>Platform level performance indicators</b>	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) can process ~3.5B calls processed per year (Max. real-time throughput: 40K transcripts/second).

For the “Accurate location prediction with high traffic and visibility” use case, we used Facebook’s state-of-the-art time series model tool Prophet<sup>16</sup> to determine the movements of users at scale, by predicting the number of users connected to a specific mobile network antenna in an urban environment. Due to the large number of time series models to train (one per antenna), we parallelized the training process with joblib, a Python library for both thread and process parallelization. We used process-based parallelization, where every CPU core was assigned a process that trained a time series model for one antenna. Overall, to train approximately 120,000 time series, it took three hours on the TID server<sup>17</sup>, which shows the model’s ability to scale. More details about the model, data, and data pre-processing is available in deliverable D3.3 [8]. The top 1000 results show mean absolute error of 1.2565, which is a low deviation from the ground truth, although we suspect that a more accurate model can be made with more data pre-processing (e.g., through the imputation of missing values). On the other hand, the model that always predicts the mean value obtained from the training set yields the mean absolute error of 11.8568 on the test set. Comparing these two models, we can notice a significant improvement in the performances (a summary of the KPIs is shown in Table 13).

**Table 13: Accurate Location Prediction with High Traffic and Visibility**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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	Predict places with high traffic and congestion event

<sup>16</sup> <https://github.com/facebook/prophet>

<sup>17</sup> 2 x Intel(R) Xeon(R) CPU E5-2695 v3 @ 2.30GHz, 56 cores, 128Gb RAM, 4 x GeForce GTX 980, 1 x 1480GB SSD Intel, 1 x 2T ST, Centos 7

			underperform at any given time; currently no automatic predictive mechanisms are in place	
<b>Application Level Performance indicators</b>	Prediction accuracy	Accuracy of forecasting as a function of time, amount of historical data, and prediction horizon	Random prediction has MAE of 11.8568	A minimum improvement of 5% (or more) over the random prediction; the I-BiDaaS solution can achieve a MAE of 1.2565 (improvement > 5% over the baseline)
<b>Platform level performance indicators</b>	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; the I-BiDaaS final solution works in a parallelized manner by harnessing the available hardware.

The use case “Optimization of Placement of Telecommunication Equipment” addresses the problem of automatizing the distribution of telecommunication equipment – more equipment is deployed for antennas that will become a hotspot (a cell site with high traffic and congestion). For this use case, we used different classification algorithms: Random Forest (implemented both in ScikitLearn and in PyCOMPSs), and two gradient boosting algorithms, namely XGBoost and CatBoost. Python’s libraries XGBoost and CatBoos, as well as the PyCOMPSs implementation of the Random Forest algorithm, are able to parallelize the training using the GPU or CPU and therefore scale to large amounts of data. The results are shown below, in Table 14.

**Table 14: Performance of binary classification algorithms for the antenna KPIs use case**

Algorithm	Accuracy	Precision	Recall
<b>XGBoost</b>	0.999	0.998	0.998
<b>CatBoost</b>	0.999	0.977	0.961
<b>Scikit-Learn Random Forest</b>	0.999	0.998	0.975
<b>PyCOMPSs Random Forest</b>	0.999	0.995	0.997

We compare our models against the model that predicts the majority class – data provided by TID for this use case has 0.012% of the positive class (i.e., an antenna becomes a hotspot), which means that the baseline model misses all the hotspots. Hence, it achieves an accuracy of 0.9988%, but with precision and recall equal to 0, as shown in Table 14.



**Table 15: Optimization of Placement of Telecommunication Equipment**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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.
<b>Application Level Performance indicators</b>	Prediction accuracy	Accuracy of forecasting as a function of time, amount of historical data, and prediction horizon	Random prediction achieves an accuracy of 0.4992 (with precision/recall equal to 0.011/0.4933 respectively) while a more competitive baseline such as the majority class classifier achieves an accuracy of 0.9988%, (with precision/recall equal to 0.0/0.0 respectively).	A minimum improvement of 5% (or more) over the random prediction; the I-BiDaaS solution achieves an accuracy of 0.999, while retaining precision and recall to 0.998 and 0.998, respectively.
<b>Platform level performance indicators</b>	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; the I-BiDaaS final solution works in a parallelized manner by harnessing the available GPU and/or CPU hardware.

### Banking experiments

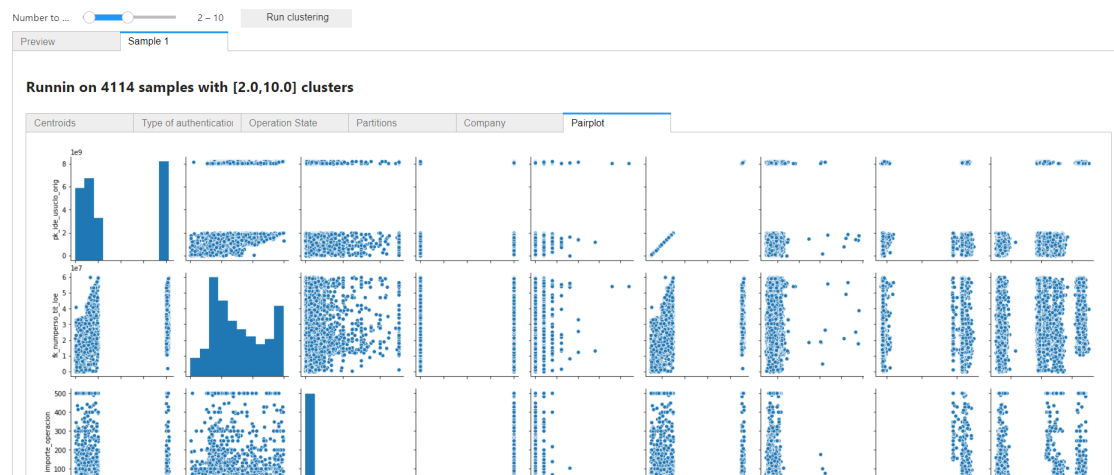
Within the “Enhance control of customers to online banking “ use case (Table 16), I-BiDaaS allowed CaixaBank’s “Intermediate users” and “Non-IT users” to modify the number of clusters and run the algorithm over a selected dataset of transactions in a very fast and easy way.

**Table 16: Enhance control of customers to online banking**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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
<b>Application Level</b>	End-to-end execution time	Anomalies detected	minutes	seconds

<b>Performance indicators</b>	Accuracy and reliability of the analytical process	Time to get analytics results.	minutes	seconds
		Data charging time	1 week	2-days
		Time to generate business rules.	Domain specific evaluation / Number of Anomalies extracted with commercial product (Data Robot) (order or 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.	
<b>Platform level performance indicators</b>	Cost	Price of technologies	Cost of commercial product licenses (e.g. DataRobot).	Depending on the selected type of license. Order of 100k€

Moreover, it supported the exploration of clients' mobile-to-mobile transaction patterns, identifying anomalies in the authentication methods and facilitated fast and visual analysis of the results in the platform, as can be seen in an example screenshot in Figure 4.



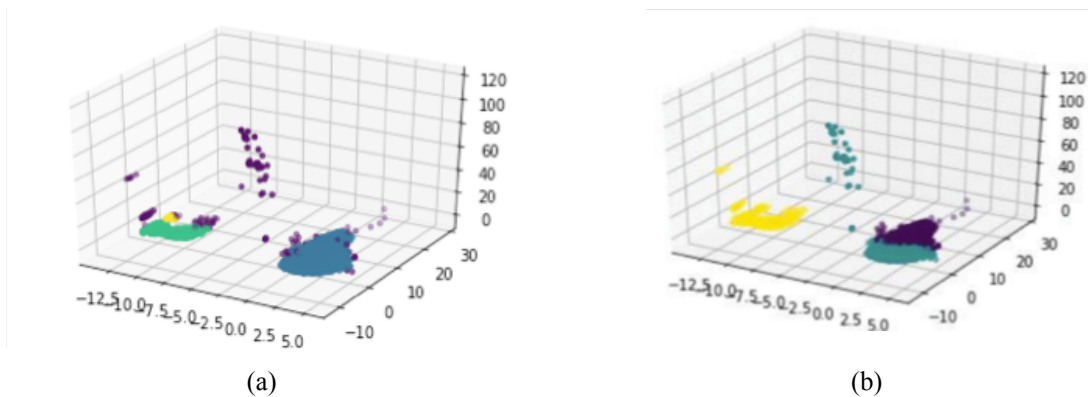
**Figure 4. Sample of the ‘Enhanced Control of customers to Online Banking’ use case clustering results in the I-BiDaaS platform**

Thanks to the I-BiDaaS solution, apart from speeding up and facilitating the access and comprehension of the data, the anomaly detection algorithm used discovered that between 5% and 10% of the bank transfers have been using an authentication mechanism while having a pattern more similar to the clusters of other authentication mechanisms (Table 17). This dataset was identified and analysed with the help of the Digital Security and Security Operation Centre (SOC) employees in order to confirm that the applied clustering algorithm was the right option for each case. This analysis was not done automatically, and still remains a manual process, but, more importantly, I-BiDaaS provided a solution that allowed to identify potential errors in our automated authentication mechanisms in mobile-to-mobile bank transfers. The obtained clusters of entries were useful to identify the different mobile-to-mobile bank transfers patterns and reconsider the way we are selecting the authentication method to proceed with the transfer.

**Table 17: Advanced Analysis of bank transfer payment in financial terminal**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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.
<b>Application Level Performance indicators</b>	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%
<b>Platform level performance indicators</b>	Cost	Price of technologies.	Cost of commercial product licenses (e.g., DataRobot).	Depending on the selected type of license. Order of 100k€

When studying the results obtained in “Advanced Analysis of bank transfer payment in financial terminal”, we can see that Figure 5(a) shows the graphical representation of the clusters generated by DBSCAN in a three-dimensional space, where the third dimension of the Principal Component Analysis (PCA) is shown as the Z-axis. The result of K-means can be examined in Figure 5(b), we can observe that some values in the Z-axis are far away from the main cluster and thus, are potential anomalies in the data.

**Figure 5. (a) DBSCAN representation, (b) K-means representation**

The PCA reduced the attributes from 501 to 3, thus, it was difficult to understand which is the correlation between the resultant three dimensions and the 501 original attributes. In Figure 6, we observe the mentioned correlation. We only show the first 84 because they are the most

interesting with respect to the 3rd dimension of the Z-axis. We also note that this third dimension is heavily influenced by attributes from 64 to 82.

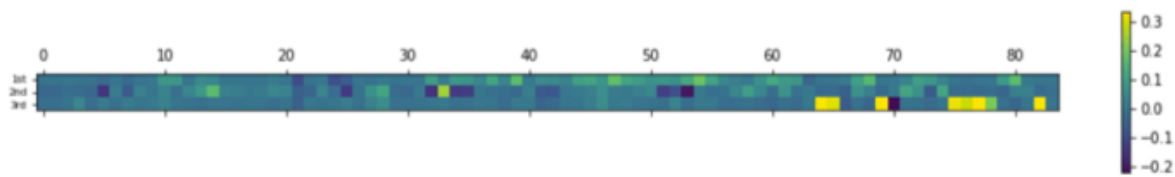


Figure 6. Heat-plot of the 84 most relevant attributes from the 501 original attributes.

I-BiDaaS not only allowed to access data and start analysing it more efficiently, but also provided a more specific subset of anomalies, along with other commercial solutions recently tested by CAIXA, such as DataRobot<sup>18</sup>. Also, I-BiDaaS allowed to define a subset of anomalies around 1-3% (depending on the dataset) in a short amount of time (just over a few seconds). DataRobot anomaly detection algorithms allowed to define a quantity of anomalies around a 5% level, but the results obtained were much less clear, being unable to clearly identify a set of anomalies and providing a much disperse dataset inside that 5%. Similarly to the “Enhance control of customers to online banking” use case, the anomalies found were sent to SOC in order to be validated manually. Approximately 70% of the reported anomalies were just uncommon operations but licit. The rest were further analysed and most of them were flagged because of internal bad practices, with just a few being actually considered as a potential fraud transaction that warranted further investigation. However, this use case demonstrated that I-BiDaaS can become a powerful tool to help SOC and Digital Security employees from CAIXA, to investigate any operation that diverges from the “normality” of the bank transfer operations undertaken at the bank offices.

The “Analysis of relationships through IP address” use case (Table 18) focused on the analysis of the synthetic data and the comparison with the real data. Results obtained from both real tokenised data and the synthetic data using those algorithms showed that the majority of the clusters found were 2-point clusters, indicating a good similarity for this use case.

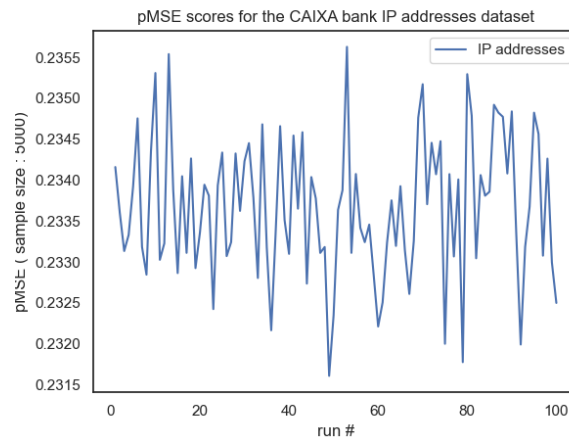
Table 18: Analysis of relationships through IP address

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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
<b>Application Level Performance indicators</b>	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)

Additional evaluation process was performed to determine a specific utility score, i.e., the similarity of results of analyses from the synthetic data and the original data. The propensity mean-squared-error (pMSE) was used as a general measure of data utility to the specific case

<sup>18</sup> <https://www.datarobot.com/>

of synthetic data. As specific utility measures, we used various types of data analyses, confidence intervals overlap and standardised difference in summary statistics, which were combined with the general utility results.



**Figure 7. Results for 100 random sampling taken from the real and the synthetic data (5K datapoints each) and the pMSE calculated using a logistic model**

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

Those quantitative results showed that the fabricated data is objectively realistic to be used for testing the use case. We further note that the rule-generation process that involves the data fabrication through TDF can be very complex and long, especially when the knowledge of the data is not complete or the extraction of rules through statistical analysis is not clear. However, once generated, new big data analysts working for CAIXA will be able to access much faster (e.g., 1-day) than going through all the data access security controls and procedures of the entity.

## Manufacturing experiments

For the “Maintenance and monitoring of production assets” use case (Table 19), CRF gathered sensor data from the production lines in order to analyse them and find a solution that may help to increase the efficiency of the machines through the monitoring of the service conditions at the device asset level. Datasets are described in Section 3.2.4 of D6.4 [4].

**Table 19: Maintenance and monitoring of production assets**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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
<b>Application Level Performance indicators</b>	Execution Time	Time to produce automated decisions	1 month	1 day
	Data Quality	Accuracy of new models with respect to internal CRF models	ND	20%
<b>Platform level performance indicators</b>	Cost	Cost regarding personnel time spent on using the system (for analysis process)	30 k€	Platform costs

		E.g. time spent for data anonymization		
--	--	--	--	--

Data were sent from CRF to the I-BiDaaS platform, where they were pre-processed and prepared for model training with an outlier detection model. Data has been transformed into separate time series – one per sensor in order to monitor each sensor separately and construct an outlier detection model for each of the time series [10]. Data stemming from the analysis are presented using a multistep approach that allows operators drill down to sensory data and detected anomalies in an intuitive and easy to use way (Figure 8). Starting from a given month (a), operators then select the category of sensors (b) they wish to see and immediately have an overview of the ones having anomalies detected. Upon selection of a sensor (c), operators see the number of the anomalies detected during the selected month (d) and can furthermore select a specific day to see the actual values and therefore review the actual anomaly that was detected (e).



**Figure 8. Sensor category selection and visualisation of anomalies**

The foundational database allows to easily visualise the history of the trend of anomalies over time and allowed CRF experts to test the efficiency of the results for the maintenance planning to avoid failures before they happen by reducing the time of decision from one (1) month to one (1) day. All of this can lead to increase the Overall Equipment effectiveness and to reduce maintenance cost.

For the “Production Process of Aluminium die-casting” use case (Table 20), CRF aggregated a large number of interconnected process parameters that influence the flow behaviour of molten metal inside a die cavity, and consequently, the productivity and the quality of the end products. Datasets are described in Section 3.2.4 of D6.4 [4].

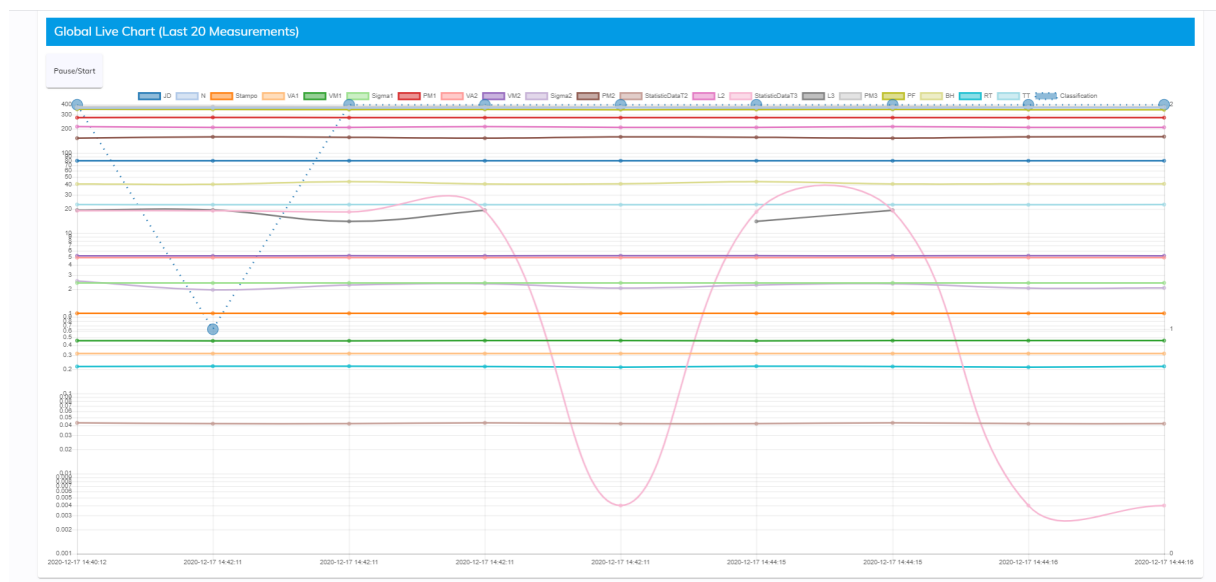
**Table 20: Production process of aluminium casting**

	Indicator	Metric	Baseline value / Benchmark	I-BiDaaS value
<b>Business KPIs</b>	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 0.5-2% (objective=1%)
<b>Application Level Performance indicators</b>	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	No baseline values. Acceptable rates are 67% of accuracy.	73%
<b>Platform level performance indicators</b>	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization	30 k€	Platform costs

CRF identified as business KPIs three (3) levels of quality: one (1) associated to high quality and two (2) associated to lower quality (1) of the engine blocks. The Advanced Machine Learning algorithm, developed within the project, allows to predict classification levels of an engine block during the die-casting process. Data are copied in real time from the CRF Server to the I-BiDaaS platform and in a few seconds, we can visualise the results of the analysis:

- i) the trend of the main important parameters (Figure 9)
- ii) a table with the identification of the engine blocks, the values of parameters and the corresponding classification level (Figure 10)
- iii) a Pie-Chart with the percentage of the classification levels and a table with the parameters measurements statistics (min, max, average) (Figure 10)



**Figure 9. Live chart with the trend of parameters for the last 20 engine blocks**



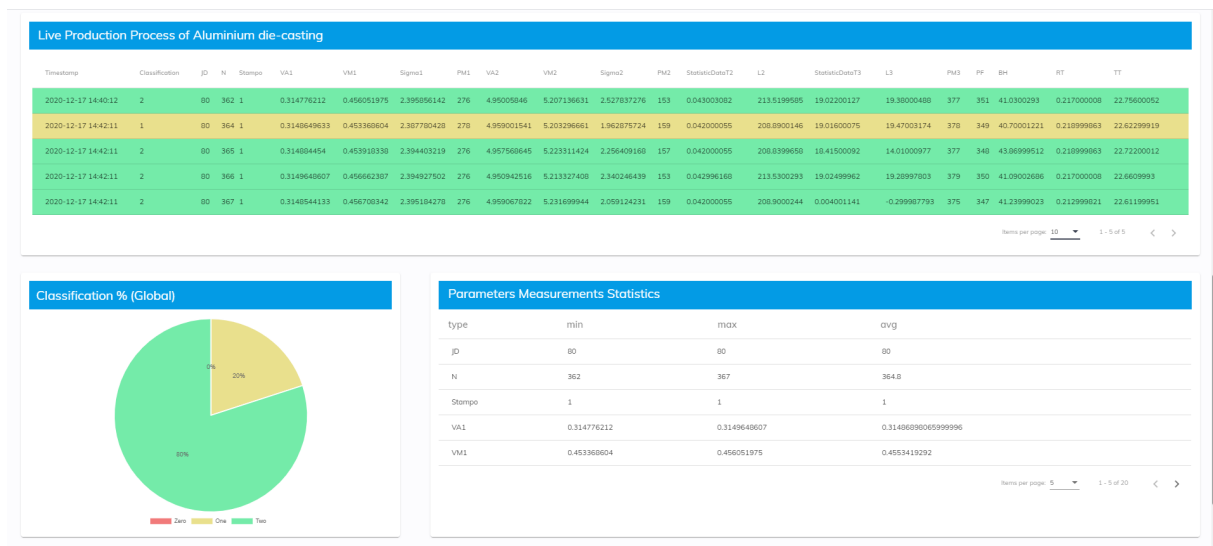


Figure 10. Colour-coded table, pie-chart and parameters statistics

The I-BiDaaS solution allows to timely visualise the classification levels of the quality of the engine blocks in the preliminary stage of the process. In this way, we can identify the sequences of the values of the parameters to increase the high quality level and decrease the lower quality levels of the engine blocks. Furthermore, by analysing the results, it is possible to reduce decision time from one month to a few hours, a turn of job or a day and avoid further processing and scraps thanks to the possibility to act before the next production steps.

### Generic experiments

For evaluating the generic use case of the I-BiDaaS platform, people were invited to perform experiments on the I-BiDaaS platform in both the Self-Service and the Expert mode. More specifically, external end-users (data analysts, big data technology providers, data consumers, etc.) were invited who either participated in previous events organised by I-BiDaaS (Info Days, Hackathons, Workshops) or are related to organisations belonging to the I-BiDaaS Consortium. Detailed walk-throughs, either for the Expert or the Self-Service mode evaluation, were prepared and shared with the end-users. The walk-throughs can be found in Annex I of this deliverable. After running the experiments, they were requested to fill in the questionnaires shown in Table 21 and Table 24, respectively.

Table 21: Questionnaire for the evaluation of the end-to-end I-BiDaaS platform in Self-Service mode

Question	Answer type	User answer
<b>How would you rate the potential of I-BiDaaS solution in the following aspects?</b> <ul style="list-style-type: none"> <li><b>Operability (real business resting possibilities)</b></li> </ul>	Score on scale 1-5 <ol style="list-style-type: none"> <li>Poor</li> <li>Average</li> <li>Good</li> <li>Very good</li> <li>Excellent</li> </ol>	
<b>How would you rate the potential of I-BiDaaS solution in the following aspects?</b> <ul style="list-style-type: none"> <li><b>Innovation (bring new features with regards to other data analytics solutions)</b></li> </ul>	Score on scale 1-5 <ol style="list-style-type: none"> <li>Poor</li> <li>Average</li> <li>Good</li> <li>Very good</li> <li>Excellent</li> </ol>	
<b>How would you rate the potential of I-BiDaaS solution in the following aspects?</b> <ul style="list-style-type: none"> <li><b>Compliance (assurance compliance with the current security and privacy regulations)</b></li> </ul>	Score on scale 1-5 <ol style="list-style-type: none"> <li>Poor</li> <li>Average</li> <li>Good</li> <li>Very good</li> </ol>	



	5. Excellent	
<b>How would you rate the potential of I-BiDaaS solution in the following aspects?</b> <ul style="list-style-type: none"> <li>Cost reduction (to what extent I-BiDaaS solution could be cost-effective compared to current solutions)</li> </ul>	Score on scale 1-5 1. Poor 2. Average 3. Good 4. Very good 5. Excellent	
<b>How satisfied are you with your I-BiDaaS experience?</b> <ul style="list-style-type: none"> <li>Overall impression</li> </ul>	Score on scale 1-5 1. Very dissatisfied 2. Somewhat dissatisfied 3. Neither satisfied nor dissatisfied 4. Somewhat satisfied 5. Very satisfied	
<b>How satisfied are you with your I-BiDaaS experience?</b> <ul style="list-style-type: none"> <li>User guidance and usability</li> </ul>	Score on scale 1-5 1. Very dissatisfied 2. Somewhat dissatisfied 3. Neither satisfied nor dissatisfied 4. Somewhat satisfied 5. Very satisfied	
<b>How satisfied are you with your I-BiDaaS experience?</b> <ul style="list-style-type: none"> <li>Information output(results)</li> </ul>	Score on scale 1-5 1. Very dissatisfied 2. Somewhat dissatisfied 3. Neither satisfied nor dissatisfied 4. Somewhat satisfied 5. Very satisfied	
<b>Did you experience any crash or malfunction while using the I-BiDaaS platform</b>	Please indicate the number of crashes/malfunctions	
<b>Number of tasks completed:</b> <ol style="list-style-type: none"> <li>Create new project,</li> <li>run experiment</li> <li>see the results</li> </ol>	Please indicate the Percentage of tasks completed (%) (please also indicate problematic tasks)	
<b>How easy and intuitive did you find using the I-BiDaaS platform when it comes to:</b> <ol style="list-style-type: none"> <li>Project Setup</li> <li>Selecting a data source</li> <li>Algorithm selection and setup</li> <li>Results Visualisation</li> </ol>	Score on scale 1-5 1. Poor 2. Average 3. Good 4. Very good 5. Excellent	<ol style="list-style-type: none"> <li>Project Setup – Score ____</li> <li>Selecting a data source – Score ____</li> <li>Algorithm selection and setup – Score ____</li> <li>Results Visualisation – Score ____</li> </ol>

Table 22: Questionnaire for the evaluation of the end-to-end I-BiDaaS platform in Expert mode

Question	Answer type	User answer
<b>Time efficiency Productivity improvement (Decrease, no change, slight increase, moderate increase, high increase)</b>	Subjective assessment of the experimental subject with regards to the I-BiDaaS solution ( <b>Decrease, no change, slight increase, moderate increase, high increase</b> with regards to previous experience)	
<b>Service quality Correspondence of I-BiDaaS platform to user needs (scale 1-5)</b>	Score on scale 1-5	

<b>Cost/Effort reduction (because of not having to maintain or setup a PyCOMPSs environment)</b>	Score on scale 1-5	
<b>Usability with respect of the tasks required:</b> 1. Create new project, 2. upload code, 3. run experiment 4. download results	score on scale 1-5 for each workflow element	<ul style="list-style-type: none"> <li>Task 1 score:</li> <li>Task 2 score:</li> <li>Task 3 score:</li> <li>Task 4 score:</li> </ul>
<b>Lines of code to develop an algorithm based on code template</b>	Number of lines	
<b>Percentage of code templates completed without problems</b>	Percentage	
<b>Number of tasks completed:</b> 4. Create new project, 5. upload code, 6. run experiment 7. download results	Percentage (please also indicate problematic tasks)	

The results of the experiments conducted by the external end-users are summarised in Table 23 (self-service mode) and Table 24 (expert mode) below. Ten (10) external end-users participated in the self-service mode evaluation and twelve (12) in the expert mode evaluation.

**Table 23: Experiment for end-to-end I-BiDaaS platform in Self-Service mode**

	Indicator	Metric	I-BiDaaS value
<b>Business KPIs</b>	Operability (real business resting possibilities) (Q4)	Evaluator's score on scale 1-5	4 (mode value)
	Innovation (bring innovative features) (Q4)	Evaluator's score on scale 1-5	4 (mode value)
	Compliance (assurance compliance with the current security and privacy regulations) (Q4)	Evaluator's score on scale 1-5	3 (mode value)
	Cost reduction (to what extent I-BiDaaS solution could be cost-effective compared to current solutions) (Q4)	Evaluator's score on scale 1-5	4 (mode value)
<b>Application Level Performance indicators</b>	Overall Impression (Q1/Q2)	Evaluator's score on scale 1-5	4 (mode value)
	User guidance and usability (Q1/Q2)	Evaluator's score on scale 1-5	5 (mode value)
	Information output (results) (Q1/Q2)	Evaluator's score on scale 1-5	5 (mode value)
	Usability with respect to each element in the workflow: (Q1/Q2) 1. Project Setup 2. Selecting a data source 3. Algorithm selection and setup 4. Results Visualisation	Evaluator's score on scale 1-5 for each workflow element	1. 5 2. 5 3. 4 4. 5 (mode values)
<b>Platform level performance indicators</b>	Correctness (Q3)	Number of crashes or malfunctions while using the platform	Zero
	Effectiveness (Q3)	Number of tasks completed / Total number of tasks	100

**Table 24: Experiment for end-to-end I-BiDaaS platform in Expert mode**

	Indicator	Metric	I-BiDaaS value
<b>Business KPIs</b>	Time efficiency	Productivity improvement (Decrease, no change, slight increase, moderate increase, high increase)	4 (mode value)
	Service quality	Correspondence of I-BiDaaS platform to user needs (scale 1-5)	5 (mode value)
	Cost/Effort reduction (because of not having to maintain or setup a PyCOMPSs environment)	Evaluator's score on scale 1-5	5 (mode value)
<b>Application Level Performance indicators</b>	Usability with respect to each element 1-8 in the workflow	Evaluator's score on scale 1-5 for each workflow element	5 (mode value)
	Coding effort	Lines of code to develop a new algorithm based on code template	8 lines extra (median value)
<b>Platform level performance indicators</b>	Correctness	Number of code templates completed without problems (expressed as percentage)	100% (mode value)
	Effectiveness	Number of tasks completed / Total number of tasks	100% (mode value)

Tables 24 and 25 above clearly demonstrate that the I-BiDaaS solution is successful with respect to the external evaluation. Namely, for the questions whose results are on scale 1-5, the baseline value of 3 is achieved at all the questions. Furthermore, the remaining questions on correctness and effectiveness have achieved the highest possible score, while the additional coding effort of 8 extra lines of code (in median) clearly corresponds to a minor additional coding effort with respect to the provided code template.

### 3.4. Deviation from expected results

The project has fully achieved its objectives and milestones for the period and has delivered significant results with immediate or potential impact. A detailed analysis of the overall impact of the I-BiDaaS project is provided in the section 5. The I-BiDaaS final integrated prototype has been successfully launched and 10 use cases have been demonstrated. The evaluation in relation to the expected project level innovation and achievements is executed as planned and described in D1.3 [1]. All I-BiDaaS achievements are closely monitored through the Key Performance Indicators (KPIs) defined by the Consortium, the status of which is reported in Table 25 and Table 26 of the current deliverable.

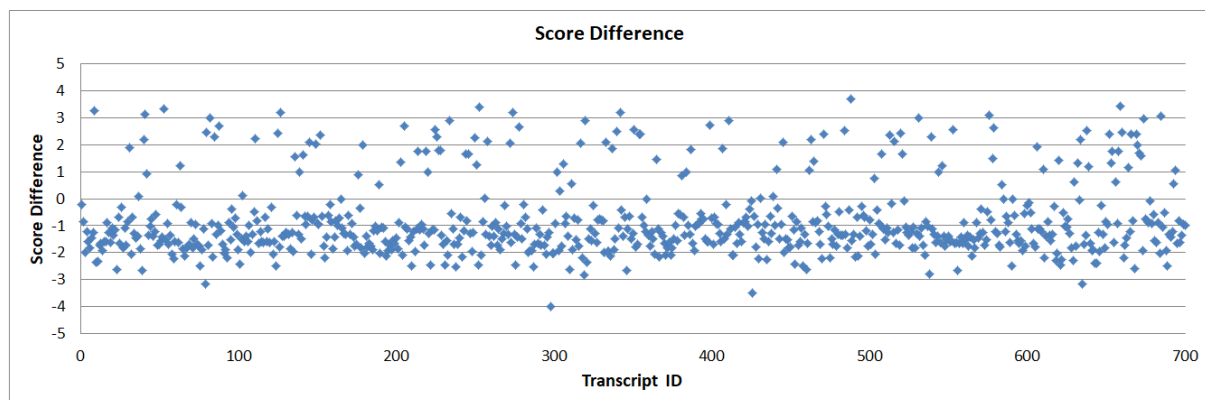
However, it is worth mentioning that during the third year of the project, the pandemic of COVID-19 affected a growing number of countries, resulting in a state of emergency, and this also applies in the case of the partner-countries of the project (mainly Italy and Spain). The social distancing policies adopted by the European governments made more challenging the work of researchers, technical and administrative staff, directly affecting the realization of project activities. An extension of 2 months for some of the project activities (i.e., Task 5.3, Task 6.1, D6.4 [4] and MS5) was requested to guarantee the achievement of all the objectives foreseen by the project. I-BiDaaS partners successfully managed to handle the delays occurred due to the pandemic.

From use cases point of view, the results have been accomplished with very few deviation from the expected results, in terms of the solution performance, flexibility and accuracy. COVID-19

impacted in the development of some cases provoking some delays on the development and deployment of use cases. For example, with respect to the TID use case on the “Quality of Service in Call Centers”, the meta-analysis we run on the correlation of the sentiment with the CSI, indicated that the sentiment score, as computed with our specific general-purpose lexicon, does not correlate with the corresponding CSI scores. However, we do not consider this a deficiency of the technology integrated within I-BiDaaS. As described in D4.3 [13], a proper domain-specific lexicon is a prerequisite to successfully predict the performance of a call center or in any other domain in general – while the reported approach is data-agnostic in that sense.

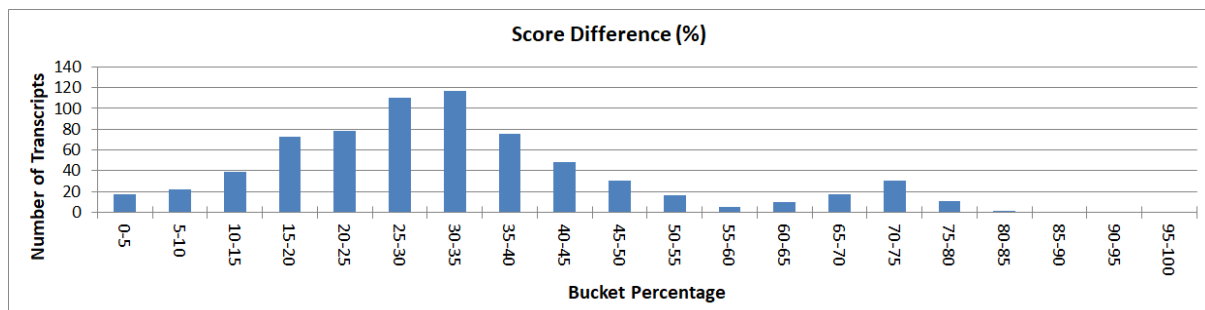
More specifically, we used as input a sample dataset of 700 transcripts (recorded in Spanish language) and a generic sentiment lexicon for Spanish terms<sup>19</sup>, both provided by the use case provider (TID). Each transcript has been marked with a CSI score that is reported by the customer at the end of the call; the CSI score ranges from 1 (very negative) to 5 (very positive). Our goal was to compare these CSI scores with the corresponding Sentiment scores that are computed by the I-BiDaaS solution for each of the transcripts, for the given sentiment lexicon. For easier comparison of sentiment and CSI scores, we normalized the sentiment scores to the same scale as CSI (i.e., [1-5]).

Figure 11 shows the actual differences of Sentiment and CSI scores (y-axis) for each of the transcripts of the input dataset (x-axis). In general, we can see that the Sentiment score is lower than the corresponding CSI score for almost all transcripts (~83%).



**Figure 11. The actual difference of Sentiment and CSI scores for each of the transcripts in our dataset.**

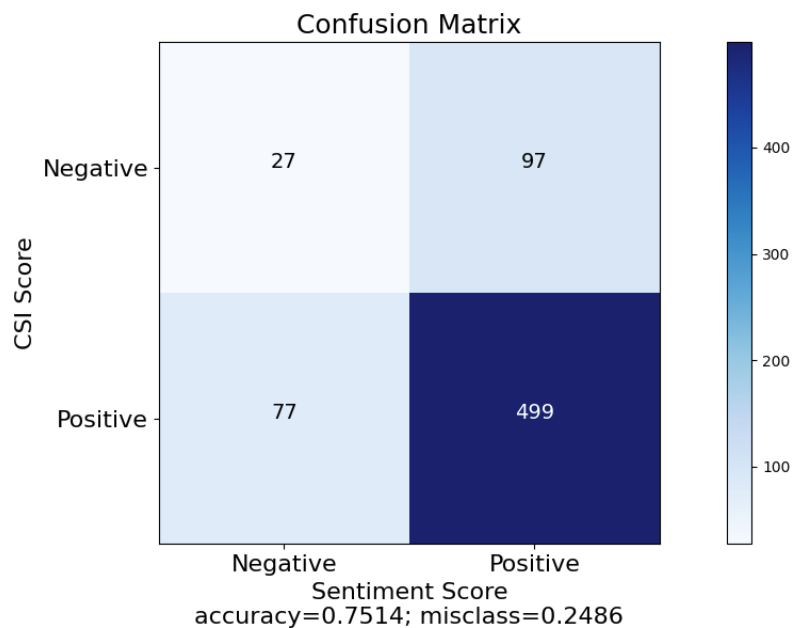
Figure 12 shows the corresponding percentage of differences accumulated across different buckets. As we can see, the majority of transcripts (~42%) differ between 20-40%, with a peak value (i.e., 118 transcripts) at around 30-35%. Only 5% of the transcripts are accumulated between 0-10%. The corresponding Pearson correlation coefficient is equal to 0.10.



**Figure 12. The accumulated number of transcripts for a given percentage difference.**

<sup>19</sup> <http://www.lsi.us.es/~fermin/index.php/Datasets> (ML-SentiCon: A Layered, Multilingual Sentiment Lexicon (English, Spanish, Catalan, Galician, Basque)).

Finally, Figure 13 presents the confusion matrix with the overall performance for binary classification. The corresponding F1-score for the positive transcripts is 0.83 and for the negative transcripts is 0.23, while the F2-score is 0.93 and 0.06, respectively.



**Figure 13. Confusion matrix of the overall performance.**

Our preliminary results indicate that the Sentiment scores computed with the specific lexicon do not correlate with the corresponding CSI scores. Further manual examination revealed that this is a consequence of the specific lexicon used, which is actually a generic sentiment lexicon for the Spanish language and is biased towards domain-general contexts. Such a lexicon cannot provide a useful indication regarding the resulting customer satisfaction levels; a domain-specific lexicon with appropriately weighted terms tailored for revealing customer satisfaction in the domain of call-centers should be used instead as input. However, because of the time restrictions due to the COVID-19, this meta-analysis of secondary importance was not pursued further. Another minor deviation was the delay of the Telefonica Research Hackathon. The said physical hackathon was originally scheduled to take place between the months of Feb-Mar, 2020, at the premises of TID in Barcelona, Spain. The COVID-19 pandemic forced us to cancel the original event. After a period of examination, we proceeded with the online version of the event, which was held out between the dates 23-25<sup>th</sup> of October.

In addition, when considering the CAIXA use cases, we were able to accomplish all the expected objectives in time. With regards to the initial expectation of the synthetic data usage, it was adapted during the project lifetime. In the beginning, CAIXA users understood that the synthetic data would allow them to deploy all the use cases and evaluate the accuracy of the platform with the synthetic data. However, some limitations were found, especially on the anomaly detection analysis of “Advanced Analysis of bank transfer payment in financial terminal” use case. Therefore, new approaches for doing the data analytics were explored, tested and evaluated, acquiring additional and very valuable results for CAIXA (such as studying the usage of different types of encryption, doing Big Data analytics over encrypted data through the I-BiDaaS platform, and extracting usable results for the Security Operations Center).

CRF had to deal with some deviations due to the shut down of most production plants and reduction of the essential work shifts that did not allow to gather new data. Specifically, for the ‘Maintenance and Monitoring of Production assets use case’, the initial intention was to use MES and SCADA data, but over time, there were problems retrieving MES data because of the scheduled activities and changes in the production lines, partially due to the COVID-19

Pandemic. From the end of May, 2020 CRF has been able to provide only new real anonymized SCADA data. This did not create any problems because analyses performed within the project showed that relevant information is in SCADA data. These data have been used to obtain thresholds for anomalous measurements for all sensors and to build a foundational database with the history of anomalies that allowed CRF to check outliers for the continuous and periodic control of the service conditions (Predictive Maintenance). For the 'Production Process of Aluminium die-casting', the data gathered before the COVID-19 Pandemic have been utilised to test the efficiency of the I-BiDaaS platform. Data has been sent every day from the CRF server to the I-BiDaaS platform by copying them every two minutes, as in the real process conditions. By simulating the real-time, CRF had the possibility to test the efficiency of the I-BiDaaS solution that provides automated results in a few seconds. CRF was able to visualise in real-time the trend of parameters and the quality classification levels of the engine blocks, calculated by the Advanced Machine Learning algorithms developed within the project.

Considering the potential deviation from expected results with reference to Subsection 5.4 of deliverable D1.3 [1] ("Evaluation from the viewpoint of subjects participating in the experiments"), there are no significant deviations with respect to the planned progress. We comment on several planned methods of engagement of participants on a case by case basis, and the activities realized within the project. Besides external evaluation, this includes the experiments that have been conducted so far by the I-BiDaaS data providers and their relevant contacts.

- Workshops: CRF has organized a hackathon at Campus Melfi<sup>20</sup> where the I-BiDaaS technologies were benchmarked on real data with respect to similar solutions, methods, and practices by the relevant actors and community in the Melfi region. See D6.2 [2] for details. In addition, TID organized a hackathon<sup>21</sup> where a number of external SME teams had an opportunity to analyze a high value TID data set. Finally, CAIXA organised a workshop<sup>22</sup> where the I-BiDaaS solution was presented to a number of external personnel from banking industry, who also provided external evaluation of the I-BiDaaS platform. The results of this evaluation are available in this deliverable.
- INFO days: I-BiDaaS has organized an Info Day<sup>23</sup> in Novi Sad, Serbia, at month 13 of the project where significant feedback from the local SME community and data science community has been received on the I-BiDaaS MVP. See D6.1 [6] and D6.2 [2] for details.
- After the release of MVP, the data providers continuously provided feedback to the technology providers during teleconferences and physical meetings (plenary and technical meetings). Further, the CAIXA workshop included an open session where external personnel could test the I-BiDaaS platform online. This possibility was also given to the EAB members at the 2020 I-BiDaaS EAB meeting. Finally, a number of external evaluators have tested in 2020 the platform in both expert and self-service modes. The evaluation results are provided in this deliverable.
- By means already put in place by the Consortium, namely TID (WAYRA), TID (AURA), and CRF (Campus-Melfi). Campus Melfi was already exploited as the location for the hackathon organized at month 16 of the project, where the Consortium exploited the possibility for onsite access to real anonymized data not available to the consortium before. TID's Aura was involved in the organization of the TID hackathon.

<sup>20</sup> <http://ibidaas.eu/blog/I-BiDaaS%E2%80%93CRF-Hackathon>

<sup>21</sup> <https://www.ibidaas.eu/blog/Telefonica-Research-Online-Hackathon/>

<sup>22</sup> <https://www.ibidaas.eu/blog/I-BiDaaS-Application-to-the-Financial-Sector-Workshop/>

<sup>23</sup> <http://ibidaas.eu/blog/I-BiDaaS-Info-Day-Workshop>



## 4. Compliance with generic big data pipelines and blueprints

An important aspect of the validation phase of big data solutions is the comparison with generic pipelines and blueprints that provide a common framework of understanding about Big Data challenges along the Big Data value chain.

The DataBench generic data pipeline<sup>24</sup> developed in the context of the DataBench project<sup>25</sup> provides an overall perspective on Big Data systems that can be specialised in order to describe more specific pipelines, depending on the type of data and the type of processing (e.g., IoT data and real-time processing). It contains 4 major steps:

- Data Acquisition / Collection (including data ingestion, processing, streaming, extraction and ingestion storage),
- Data Preparation / Storage (including storage retrieval / access / queries, data protection, curation, integration and publication),
- Data Analytics (including data processing for analysis, AI and Machine Learning)
- Data Visualization / Interaction (including data presentation, environment / boundary / user action and interaction).

These steps are in compliance with the activities described in the Reference Architecture for Big Data Application of the Big Data Value Association (BDVA)<sup>26</sup>.

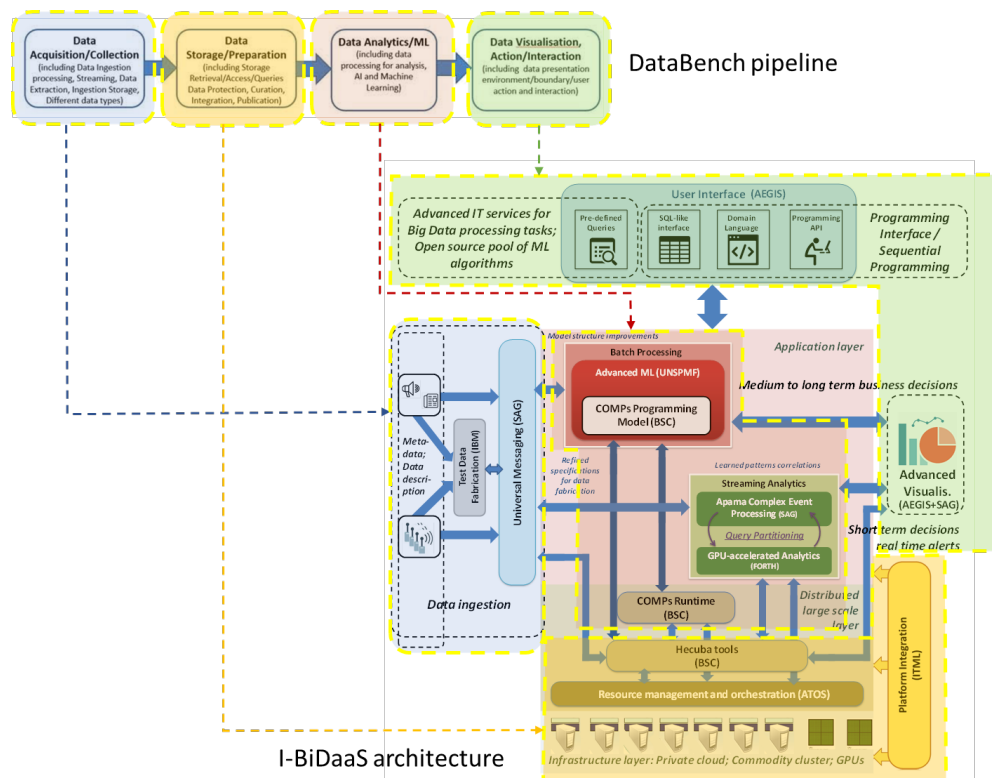


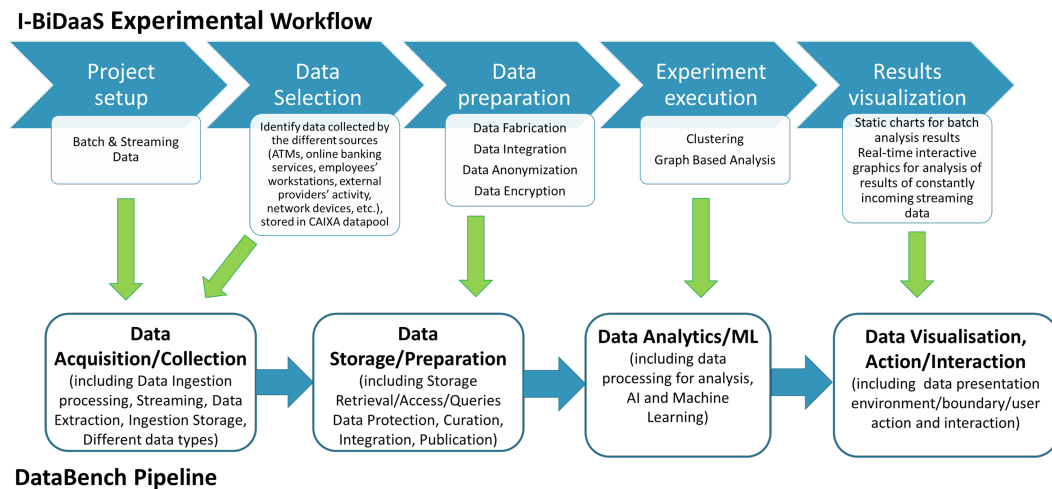
Figure 14. Mapping of I-BiDaaS architecture to the DataBench pipeline

<sup>24</sup> DataBench Consortium, D4.4 DataBench Benchmarking Handbook, Project Deliverable, 31/10/2020 available at [https://www.databench.eu/wp-content/uploads/2020/11/databench-d.4.4\\_1.0.pdf](https://www.databench.eu/wp-content/uploads/2020/11/databench-d.4.4_1.0.pdf)

<sup>25</sup> <https://www.databench.eu>

<sup>26</sup> Big Data Value Strategic Research and Innovation Agenda (BDV SRIA). [http://bdva.eu/sites/default/files/BDVA\\_SRIA\\_v4\\_Ed1.1.pdf](http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf)

The I-BiDaaS solution is fully aligned with the above pipeline both in terms of the system architecture (described in D5.6 [5]) as well as in terms of the I-BiDaaS experimental workflow (described in D6.4 [4]), as shown in Figure 14 and Figure 15, respectively.



**Figure 15. Alignment between I-BiDaaS banking experimental workflow and the DataBench pipeline in the context of the banking experiments**

This alignment enables the mapping of the I-BiDaaS solution to the generic Big Data Analytics Blueprint (devised in DataBench), as well as the BDVA Reference Architecture for Big Data Applications. The advantages of this mapping are twofold. First, it enables the positioning of the I-BiDaaS solution and its components in the context of related Big Data technologies along the Big Data value chain and secondly, it facilitates the exploitation of the industrial benchmarks and knowledge developed in the context of DataBench in order to compare the experimental results obtained in I-BiDaaS with those obtained in the context of similar business cases.

DataBench benchmarks have been calculated collecting data from the relevant European industries in which Big Data have the highest impacts and represent the average improvement achieved. They correspond to quantitative KPIs, such as cost reduction, as well as qualitative KPIs such as time efficiency and product/service quality. Such KPIs have also been used in the context of the I-BiDaaS experiments.

For example, the benchmark value for time efficiency in the financial sector according to DataBench calculations is 3, corresponding to an improvement range of 10–24%. Time efficiency is also one of the KPIs used in the CAIXA experiment “Advanced Analysis of bank transfer payment in financial terminal”. The I-BiDaaS value reported in section 3.3, Table 18 (reduction of time to access data from two weeks to two days) represents an improvement of approximately 600%, which is significantly higher than the median value for this industry.

Finally, the alignment of the I-BiDaaS solution with the DataBench generic pipeline enables the inclusion of the I-BiDaaS use cases to the DataBench ToolBox, thus contributing to the development of European empirical evidence of the business benefits that can be achieved using Big Data Technologies.



## 5. Impact Analysis

### 5.1. Industrial impact assessment

As stated in the Description of Action, the I-BiDaaS project was expected to provide services and tools to enhance big data processing performance in tasks both for non-IT users and for IT-users/developers. The motivation behind the targeted call was to develop technologies that would increase the efficiency of all EU companies and organisations that need to manage vast and complex amounts of data and in particular, the competitiveness of EU enterprises. The emphasis was on rigorously measured increases in performance in data processing at a very large scale. That led to the formalisation of the I-BiDaaS value proposition.

**Value proposition: I-BiDaaS** will deliver a full array of big data business analytics solutions for structured, unstructured, noisy and potentially synthetic data for companies in multiple industries (finance, telecom and automotive) that are more accessible, cost effective and employee-empowering than existing solutions, which gives companies the confidence 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.

The key takeaways of the I-BiDaaS project were to (i) deliver all expected impacts as stated in the call, (ii) improve data exchange, storing and interoperability standards, and (iii) foster innovation in terms of processing tools and methods that are applicable in real-world settings and increase speed of data throughput and access. Moreover, to ensure that the I-BiDaaS project meets its ambitious objectives, and achieve expected impacts, a three-stage impact assessment model was used 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 Consolidation Phase (M33-M36) of the project, along with an overall summary with respect to the expected project impact level, are discussed. Moreover, the implemented activities aiming to demonstrate the I-BiDaaS solution and involve external users in the evaluation process are described.

#### I-BiDaaS Impact and measures used

In I-BiDaaS, specific ambitious objectives were defined from the very beginning of the project, aiming to increase the impact in the research community, in the Data Market and the Big Data EU Economy and also to contribute to standards, initiatives and to innovation capacity. All I-BiDaaS achievements towards the delivery of the expected impact were closely monitored through the Key Performance Indicators (KPIs) defined by the consortium. They are an indispensable management tool that allowed us to monitor the progress, to enable evidence-based decision-making, and to aid in the development of strategies. In the following tables (Table 25 and Table 26), the overall progress with respect to the KPIs is reported.

**Table 25: I-BiDaaS KPIs**

What the call states		
<b>KPI-RI-1</b>	Release of I-BiDaaS framework and tools under an open-source non-viral license.	<b>Open source version of the I-BiDaaS platform including 5 components published</b>

		4 tools are provided as open source
<b>KPI-RI-2</b>	Increased speed of data analysis and throughput (compared to industrial-based benchmarks) by more than 10%.	See Table 26 below. For each use case provider, the related numbers are reported.
<b>KPI-RI-3</b>	Increase in the direct access of big data analytics tools by more than 30%.	<b>100%</b>
<b>KPI-RI-4</b>	Define at least 2 standards related to Big Data Analytics and uptake at least 5.	See Table 26 below. Mapping between requirements, activities and functional components and detailed lists with standards I-BiDaaS uptakes
<b>KPI-RI-5</b>	Influence at least 4 formal specifications of standards.	<b>4</b>
<b>KPI-RI-6</b>	Implementation of 3 data practitioners' demonstrators validating at least 80% of tools.	<b>8</b>
<b>Impact &amp; Exploitation KPIs</b>		
<b>KPI-IE-1</b>	At least 3 I-BiDaaS tools reach market readiness level at the end of the project.	2 I-BiDaaS tools reached market readiness level, another tool is almost there
<b>KPI-IE-2</b>	At least 4 standalone tools and methods delivered.	<b>5</b>
<b>KPI-IE-3</b>	At least 6 third-party collaborations to be established for further applicability verification.	<b>&gt;6</b>
<b>KPI-IE-4</b>	At least 3 experiments demonstrating the tools' applicability within I-BiDaaS.	<b>10</b>
<b>KPI-IE-5</b>	Increased programmability for users by at least 30% compared to today, verified on at least 1 practitioner.	<b>~57% less LoCs</b> using Hecuba & COMPSs for distributed programming & DB access
<b>KPI-IE-6</b>	Reduction of practitioners LOCs (lines of code) by 50% due to the ability to transform a sequential application into a parallel and distributed one.	<b>~50%</b>
<b>KPI-IE-7</b>	At least 1500 downloads of the tools through the project.	<b>1786</b>
<b>Dissemination &amp; Communication KPIs</b>		
<b>KPI-DC-1</b>	At least 500 downloads for public deliverables, prototypes, promotional material.	<b>2806 Direct Downloads</b>
<b>KPI-DC-2</b>	At least 10 publications.	<b>3 Journal, 18 Conference and Workshop Papers, 4 Posters</b>
<b>KPI-DC-3</b>	At least 3 conferences or workshop participations per year.	<b>2018: 7 Conf. &amp; 4 Workshops 2019: 9 Conf. &amp; 2 Workshops 2020: 8 Conf. &amp; 3 Workshops</b>
<b>KPI-DC-4</b>	At least 33% of conference and journal papers have an impact factor or ERA classification.	<b>50%</b>
<b>KPI-DC-5</b>	At least 33% gold open-access journal articles.	<b>33.3%</b>
<b>KPI-DC-6</b>	At least 2 active participations to a standardization body.	<b>2</b>
<b>KPI-DC-7</b>	At least 2 standards that are used and improved within I-BiDaaS.	Detailed lists with standards for the use cases and the components have been identified.
<b>KPI-DC-8</b>	At least 3 workshops or special events.	<b>9</b>
<b>KPI-DC-9</b>	At least 3 collaborations with projects in H2020.	<b>5</b>
<b>KPI-DC-10</b>	At least 4 participations to collaborative initiatives.	<b>6</b>

Table 26: Explanation of the I-BiDaaS KPIs

<b>What the call states</b>
-----------------------------

KPI-RI-1	<p>One of the goals of the I-BiDaaS consortium is to contribute to the Open Source communities to provide benefit to the European community. COMPSs 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.</p> <p>Finally, to support the sustainability of the I-BiDaaS technical achievements and disseminate the expertise that the consortium gathered during the project, it was decided to create and publish an open-source version of the I-BiDaaS platform. The open-source version can be used by the community as a Big Data analytics platform. The components from the I-BiDaaS platform that are included in the open-source version are:</p> <ul style="list-style-type: none"><li>• The Advanced Visualisation Toolkit (AVT) by AEGIS</li><li>• The Orchestrator by ITML</li><li>• The COMPSs/PyCOMPSs programming model and the Hecuba Tools by BSC</li><li>• The docker image<sup>27</sup> that contains code and libraries developed by UNSPMF and BSC is used for running the batch processing experiments</li><li>• The APAMA community version by SAG. In order to create and run a streaming use case in the open-source version, UM (that is a commercial product) can be replaced by an open-source MQTT compatible broker like Mosquitto</li></ul> <p>The installer is available in the I-BiDaaS knowledge repository<sup>28</sup>.</p>																										
KPI-RI-2	<table><tr><th>CRF</th><th>Before I-BiDaaS</th><th>With I-BiDaaS</th></tr><tr><td rowspan="3"></td><td>In the real industrial scenario the time to analyse data and make decisions is 1 month for both use cases.</td><td><i>Aluminium die-casting use case:</i> data came to the platform every two minutes and in a <b>few seconds</b> we can <b>visualise the results</b> (Automated analysis with <b>response in real time</b>). The decision time is reduced to a few hours, a turn of job or 1 day depending on the type of response. <i>Maintenance and monitoring of production assets: 1 day.</i></td></tr><tr><td>Manual statistical analysis, R2 low, with error of prediction &gt; 30%</td><td>Improved classification accuracy, <b>73%</b></td></tr><tr><td>Standard analysis method, with no update.</td><td><b>New</b> potential insights.</td></tr><tr><th>TID</th><th>Before I-BiDaaS</th><th>With I-BiDaaS</th></tr><tr><td rowspan="2"></td><td>~<b>11,520</b> calls processed (per year) by a single human agent.</td><td>~<b>3.5B</b> calls processed per year in a single GPU; max. real time throughput: 40K transcripts/second.</td></tr><tr><td>The manual process allows to identify about <b>2,300</b> low customer satisfaction audio calls.</td><td>Increase the number of detected low customer satisfaction calls by a human agent to <b>7,000 (200% increase)</b> by ranking the calls based on sentiment.</td></tr><tr><th>CAIXA</th><th>Before I-BiDaaS</th><th>With I-BiDaaS</th></tr><tr><td rowspan="2"></td><td><b>2 weeks</b> to provide data to a new data analytics provider.</td><td><b>One day</b> to get data ready for analysis.</td></tr><tr><td><b>Limited</b> insights and anomalies found.</td><td><b>New</b> insights found ; validated @CAIXA.</td></tr></table>	CRF	Before I-BiDaaS	With I-BiDaaS		In the real industrial scenario the time to analyse data and make decisions is 1 month for both use cases.	<i>Aluminium die-casting use case:</i> data came to the platform every two minutes and in a <b>few seconds</b> we can <b>visualise the results</b> (Automated analysis with <b>response in real time</b> ). The decision time is reduced to a few hours, a turn of job or 1 day depending on the type of response. <i>Maintenance and monitoring of production assets: 1 day.</i>	Manual statistical analysis, R2 low, with error of prediction > 30%	Improved classification accuracy, <b>73%</b>	Standard analysis method, with no update.	<b>New</b> potential insights.	TID	Before I-BiDaaS	With I-BiDaaS		~ <b>11,520</b> calls processed (per year) by a single human agent.	~ <b>3.5B</b> calls processed per year in a single GPU; max. real time throughput: 40K transcripts/second.	The manual process allows to identify about <b>2,300</b> low customer satisfaction audio calls.	Increase the number of detected low customer satisfaction calls by a human agent to <b>7,000 (200% increase)</b> by ranking the calls based on sentiment.	CAIXA	Before I-BiDaaS	With I-BiDaaS		<b>2 weeks</b> to provide data to a new data analytics provider.	<b>One day</b> to get data ready for analysis.	<b>Limited</b> insights and anomalies found.	<b>New</b> insights found ; validated @CAIXA.
CRF	Before I-BiDaaS	With I-BiDaaS																									
	In the real industrial scenario the time to analyse data and make decisions is 1 month for both use cases.	<i>Aluminium die-casting use case:</i> data came to the platform every two minutes and in a <b>few seconds</b> we can <b>visualise the results</b> (Automated analysis with <b>response in real time</b> ). The decision time is reduced to a few hours, a turn of job or 1 day depending on the type of response. <i>Maintenance and monitoring of production assets: 1 day.</i>																									
	Manual statistical analysis, R2 low, with error of prediction > 30%	Improved classification accuracy, <b>73%</b>																									
	Standard analysis method, with no update.	<b>New</b> potential insights.																									
TID	Before I-BiDaaS	With I-BiDaaS																									
	~ <b>11,520</b> calls processed (per year) by a single human agent.	~ <b>3.5B</b> calls processed per year in a single GPU; max. real time throughput: 40K transcripts/second.																									
	The manual process allows to identify about <b>2,300</b> low customer satisfaction audio calls.	Increase the number of detected low customer satisfaction calls by a human agent to <b>7,000 (200% increase)</b> by ranking the calls based on sentiment.																									
CAIXA	Before I-BiDaaS	With I-BiDaaS																									
	<b>2 weeks</b> to provide data to a new data analytics provider.	<b>One day</b> to get data ready for analysis.																									
	<b>Limited</b> insights and anomalies found.	<b>New</b> insights found ; validated @CAIXA.																									
KPI-RI-3	<p>The I-BiDaaS data providers (thanks to IBM’s TDF) created new synthetic datasets that were not available before. In addition, CAIXA, TID and CRF created within I-BiDaaS new anonymised/tokenised datasets. These synthetic and real datasets 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 tokenised and synthetic data to the I-</p>																										

<sup>27</sup> <https://hub.docker.com/repository/docker/vchatzi/ibidaas-universal>

<sup>28</sup> [https://github.com/ibidaas/knowledge\\_repository/tree/master/tools\\_technologies/ibidaas\\_installer/ibidaas\\_installer](https://github.com/ibidaas/knowledge_repository/tree/master/tools_technologies/ibidaas_installer/ibidaas_installer)

	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.
<b>KPI-RI-4</b>	<p>Contributing to a standard via BDVA.</p> <p>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.</p> <p>On 12/2018, after the project proposal was submitted, a new standard, Recommendation ITU-T Y.3519<sup>29</sup>, “Cloud computing – Functional architecture of big data as a service” was released. The Recommendation describes the functional architecture for Big Data as a Service (BDaaS). The functional architecture is defined on the basis of the analysis of requirements and activities of cloud computing-based big data described in Recommendation ITU-T Y.3600.</p> <p>Following the methodology of Recommendation ITU-T Y.3502, the BDaaS functional architecture is described from a set of functional components and cross-cutting aspects. The specified functional components consist of sets of functions that are required to perform the BDaaS activities for the roles and sub-roles described in Recommendation ITU-T Y.3600.</p> <p>This standard seems to fully cover the I-BiDaaS functional architecture. For complete analysis see Table 39 in Annex II of this deliverable. The table summarizes the mapping between requirements, activities and functional components.</p> <p><b>KPI-DC-7 below</b> includes a list of standards I-BiDaaS uptakes.</p>
<b>KPI-RI-5</b>	<p>Four (4) members of I-BiDaaS (ATOS, IBM, SAG, BSC) are founding members of the Big Data Value Association (BDVA) from the European Public Private Partnership on Big Data Value (PPP BDVA). Thus, the project has been aligned with PPP activities like the Strategic Research and Innovation Agenda (SRIA). Strong collaborations were pursued with other activities of the PPP that were directly related to the “data processing architectures”, the “data analytics” and the “data visualisation and user interaction”.</p> <p>The I-BiDaaS components closely match the functional components of the Big Data Interoperability Framework of NIST.</p> <p>I-BiDaaS partners IBM and SAG are both active members of the Data Mining Group, which hosts the working group that develops and improves the Predictive Model Markup Language (PMML), a standard that simplifies the deployment of analytic models.</p> <p>The leaders of Task Force 6 (Technical) subgroup 6 (TF6.SG6) on Standardisation have organised a call in order to engage directly with BDVA members willing to be more involved in the different activities of the sub-group, including providing contents for the SRIDA update.</p> <p>The proposal to be inserted in SRIDA - ISO/IEC JTC1 SC42 has the work item ISO/IEC 23894 Information Technology — Artificial Intelligence — Risk Management</p> <p>Define AI specific requirements and guidelines for Risk Management.</p>
<b>KPI-RI-6</b>	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.
<b>Impact &amp; Exploitation KPIs</b>	
<b>KPI-IE-1</b>	<p>Qbeast by BSC has reached market readiness level and commercialization is expected to start in 2021. 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. The company creation started in 2018, taking early steps to protect the related IP with a patent application and surveying the market. In 2019, two success stories were developed, and the company received an award as the best spin-off with high technological impact. In 2020, fundraising and product development have been the main goal, while the company has been officially incorporated. The first customers are expected during 2021.</p> <p>The AEGIS Advanced Visualization Toolkit (AVT) consists of a set of interactive visualisation tools developed to allow for a more straightforward exploratory analysis. The selection of the tools and, subsequently, the definition of the interactions among them depend on the domain and on the end-users’ requirements. AVT development has started in the context of I-BiDaaS and has been developed along the project. It has reached the TRL level 6. An internal dedicated team has been formed to move the tool at a TRL7/8 level, also taking into account progress performed in the context of other EU-funded projects that AEGIS is currently participating. Finally, a license scheme together with a solid process for production deployment (including deployment paths for clients in one of the I-BiDaaS domains but also cases in entirely different domains) are being currently established by AEGIS in view of the commercialization of AVT.</p>

<sup>29</sup> [https://www.itu.int/rec/dologin\\_pub.asp?lang=e&id=T-REC-Y.3519-201812-I!!PDF-E&type=items](https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.3519-201812-I!!PDF-E&type=items)

	The IBM Test Data Fabrication (TDF) platform enables the creation of synthetic realistic data to facilitate development and testing of big data applications where real data is not available or cannot be used due to privacy/confidentiality restrictions. The TDF technology has been significantly enhanced during the I-BiDaaS project to fully support all the project use cases. The major features developed in the tool are a new UI that enables graphical form based definition of data fabrication rules and parallel CSP solver that significantly speeds up the performance of the data fabrication process and supports big-data use cases. TDF has reached market readiness level.
<b>KPI-IE-2</b>	I-BiDaaS delivers several standalone tools. For example, IBM's TDF, AEGIS' Advanced visualization toolkit, BSC's Hecuba, BSC's Qbeast and UNSPMF's machine learning algorithm implementations in COMPSs. Each of the mentioned tools is either fully created within I-BiDaaS or is significantly improved within I-BiDaaS.
<b>KPI-IE-3</b>	Third-party collaborations are focusing on SMEs and have already been established. During the course of the project, the I-BiDaaS Consortium organised nine (9) special events such as Workshops, Info Days, Hackathons, Webinars and participated to even more events (BDV PPP Summit 2019, EBDVF 2019, BDV PPP Summit 2020, EBDVF2020) 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 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 [14].
<b>KPI-IE-4</b>	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. 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.
<b>KPI-IE-5</b>	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 COMPSs and Hecuba take care of it. One could take as a simple example the following <a href="https://github.com/ibidaas/knowledge_repository/tree/master/tools_technologies/compare_LOC">https://github.com/ibidaas/knowledge_repository/tree/master/tools_technologies/compare_LOC</a> and programm it with / without COMPSs + Hecuba. If we account for meaningful code only, (skipping definitions) it is 73 lines of code vs 171, which is a ~57% reduction.
<b>KPI-IE-6</b>	As an example of how this KPI is reached, we compare an MPI <sup>30</sup> and a COMPSs implementation <sup>31</sup> for the ADMM-Lasso algorithm for sparse regression. The MPI (C-based) implementation has 216 lines of code, while the COMPSs (Python-based) implementation has 109 lines, which yields approximately a 50% reduction <sup>32</sup> .
<b>KPI-IE-7</b>	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 1786. The KPI threshold has been reached.
<b>Dissemination &amp; Communication KPIs</b>	

<sup>30</sup> <https://web.stanford.edu/~boyd/papers/admm/mpi/>

<sup>31</sup> [https://github.com/ibidaas/knowledge\\_repository/blob/master/tools\\_technologies/sources/batch\\_processing/uns\\_pmf/distributed\\_Lasso\\_ADMM.py](https://github.com/ibidaas/knowledge_repository/blob/master/tools_technologies/sources/batch_processing/uns_pmf/distributed_Lasso_ADMM.py)

<sup>32</sup> The programming languages are different (C-based MPI versus Python based COMPSs), but the comparison is reasonable, as a lower-level language (C) yields more code for a tighter resource control, while with COMPSs we have fewer lines of code at a price of moderately reduced performance; see Table 21, D6.4 [4], Advanced ML submodule.



<b>KPI-DC-1</b>	The project boosted the dissemination efforts within the duration of the project achieving 2806 downloads of its material.
<b>KPI-DC-2</b>	I-BiDaaS is delivering a solid publications list, including 3 Journal papers, 18 Conference/Workshop papers, and 4 Conference/Workshop Posters. The detailed list for 2018 publications is reported in D7.3 [15], for 2019 in D7.5 [14] and for 2020 in D7.7 [16]. I-BiDaaS publications can be found on the project's website <sup>33</sup> and also in Zenodo <sup>34</sup> and OpenAIRE <sup>35</sup> .
<b>KPI-DC-3</b>	<p>During the reporting period, I-BiDaaS partners attended 24 Conferences &amp; 9 Workshops</p> <ul style="list-style-type: none"> <li>• <b>2018:</b> 7 Conferences &amp; 4 Workshops</li> <li>• <b>2019:</b> 9 Conferences &amp; 2 Workshops</li> <li>• <b>2020:</b> 8 Conferences &amp; 3 Workshops</li> </ul> <p>The detailed list of conferences &amp; workshops attended by I-BiDaaS partners can be found in D7.3 [15] for 2018 and D7.5 [14] for 2019 and in D7.7 [16] for 2020.</p> <p>I-BiDaaS Consortium invested 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>
<b>KPI-DC-4</b>	We have achieved 50% of the publications to have an impact factor or ERA classification.
<b>KPI-DC-5</b>	<p>3 (three) journal articles have been accepted and published.</p> <ol style="list-style-type: none"> <li>1. <b>Gold</b> Open Access - Sahu, A.K., Jakovetic, D., Bajovic, D. and Kar, S., 2018. Communication efficient distributed weighted non-linear least-squares estimation. <i>EURASIP Journal on Advances in Signal Processing</i>, 2018(1), p.66.</li> <li>2. <b>Green</b> Open Access - Jakovetić, D., Krejić, N. and Jerinkić, N.K., 2019. Exact spectral-like gradient method for distributed optimization. <i>Computational Optimization and Applications</i> 74, 703-728.</li> <li>3. <b>Green</b> Open Access - Jerinkić, N.K., Jakovetić, D., Krejić, N. and Bajović, D., 2020. Distributed second-order methods with increasing number of working nodes. <i>IEEE Transactions on Automatic Control</i>, vol. 65, no. 2, pp. 846-853.</li> </ol> <p>Thus, I-BiDaaS consortium has achieved 33.3% gold open access to the journal articles linked to I-BiDaaS scientific results.</p>
<b>KPI-DC-6</b>	<ol style="list-style-type: none"> <li>1. Participation in <b>BDVA</b> which is driving big data standardization and interoperability priorities and is connected with Big Data Standards related to Big Data PPP projects.</li> </ol> <p>The I-BiDaaS solution can be contextualised within the BDV reference model defined in the BDV Strategic Research and Innovation Agenda (BDV SRIA) and contributes to the model.</p> <p>Specifically, the work is relevant to the following BDV reference model horizontal concerns:</p> <ul style="list-style-type: none"> <li>• Data visualisation and user interaction: We develop several advanced and interactive visualisation solutions applicable in the targeted sectors.</li> <li>• Data analytics: We develop data analytics solutions for the eight industrial use cases. While the solutions may not correspond to state-of-the art advances in algorithm development, they clearly contribute to revealing novel insights into how Big Data analytics can improve banking, telecom and manufacturing operations.</li> <li>• Data processing architectures: We develop an architecture that is well-suited for applications where both batch analytics (e.g., analysing historical data) and streaming analytics (e.g., online processing of new transactions) are required. A novelty of the architecture is the incorporation of realistic synthetic data fabrication and the definition of scenarios of usefulness and quality assurance of the corresponding synthetic data.</li> <li>• Data protection: We describe how data tokenisation and realistic synthetic data fabrication can be used in relevant applications to allow for more agile development of Big Data analytics solutions.</li> <li>• Data management: We present innovative ways for data management utilising efficient multidimensional indexing.</li> </ul> <p>Regarding the BDV reference model vertical concerns, the work is relevant to the following:</p> <ul style="list-style-type: none"> <li>• Big Data Types and Semantics: The work is mostly concerned with structured data, meta-data, and graph data. The work contributes to generation of realistic synthetic data from the corresponding domain-defined meta-data.</li> <li>• Cybersecurity: The presented solutions that include data tokenisation correspond to novel best practice examples for securely sharing sensitive banking data outside bank premises.</li> </ul>

<sup>33</sup> <http://www.ibidaas.eu>

<sup>34</sup> <https://zenodo.org>

<sup>35</sup> [https://explore.openaire.eu/search/project?projectId=corda\\_h2020::652e6b81a75292294cdd34ff5a806573](https://explore.openaire.eu/search/project?projectId=corda_h2020::652e6b81a75292294cdd34ff5a806573)

	<p>Therefore, in relation with BDV SRIA, we contribute to the following technical priorities: Data protection; Data Processing Architectures; Data Analytics; and Data Visualisation and User Interaction.</p> <ol style="list-style-type: none"> <li>2. Participation and active collaboration with <b>DataBench</b>, who is designing performance benchmarking processes for Big Data. DataBench is expected to set the standards and benchmarks for the emerging Big Data ecosystem. <ul style="list-style-type: none"> <li>• The I-BiDaaS data providers (CAIXA, CRF, TID) participated in a survey carried out by DataBench on Big Data use and impacts and responded to a questionnaire released by DataBench to check the validity of their benchmarks.</li> <li>• On July 7, 2020, I-BiDaaS participated in a webinar organized by the DataBench project entitled “Virtual BenchLearning – Assessing the Performance and Impact of Big Data, Analytics and AI”. 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 in the webinar through a presentation of the current I-BiDaaS benchmarking approach, landscape and needs, both from the technological and business perspectives.</li> <li>• On November 4, 2020, I-BiDaaS participated at the DataBench Final Event in the framework of EBDVF2020. We delivered a general presentation of I-BiDaaS, as well as information on how the I-BiDaaS experimentation maps to the generic Big Data pipelines described in DataBench. In addition, I-BiDaaS and DataBench had several online meetings that facilitated collaboration.</li> <li>• An important dimension of this collaboration is that DataBench provided to I-BiDaaS several recommendations on concrete benchmarks that are applicable to the I-BiDaaS main targeted sectors—manufacturing, telecommunication, and banking.</li> <li>• I-BiDaaS contributions to the DataBench project's ReachOut campaign – the campaign on Generation of architectural Pipelines-Blueprints (reachout-project.eu). Specifically, we report on the Mapping of I-BiDaaS architecture to the DataBench pipeline, and the Mapping of I-BiDaaS “Advanced Analysis of bank transfer payment in financial terminal” case to the DataBench generic blueprint.</li> </ul> </li> </ol>
<b>KPI-DC-7</b>	<p>I-BiDaaS, among many additional standards, references NIST and BDVA big data reference models. Among the standards and Bodies of Knowledge referenced are: BABOK, CMMI, IEEE standards, ISO 9001, ISO/IEC standards, PMBOK, SWEBOK, ITIL</p> <p>In Annex II of this deliverable, several tables providing detailed representation of the standards used in the various use cases and components of I-BiDaaS are included. In more detail:</p> <ul style="list-style-type: none"> <li>• Table 34: List of Standards used in I-BiDaaS Technologies</li> <li>• Table 35: List of Standards used in the use-case experiments</li> <li>• Table 36: List of Standards per layer related to I-BiDaaS</li> <li>• Table 37: Analysis of standard processes and methodologies per use-case</li> <li>• Table 38: Participation to Standardisation Bodies per partner</li> </ul>
<b>KPI-DC-8</b>	<p>Nine (9) special events have been organized in the context of I-BiDaaS:</p> <ol style="list-style-type: none"> <li>1. <b>I-BiDaaS Info Day - Workshop on Big Data Analytics</b><sup>36</sup>: January 22, 2019, Faculty of Sciences of University of Novi Sad, Serbia</li> <li>2. <b>CRF's hackathon at Campus Melfi</b><sup>37</sup>: June 18-19, 2019, Campus Melfi, Italy.</li> <li>3. <b>Satellite Promotional Event at BDV PPP Summit in Riga</b><sup>38</sup>: June 26-28, 2019, Riga, Latvia</li> <li>4. <b>European Big Data Value Forum 2019</b><sup>39</sup>: October 14-16, 2019, Helsinki, Finland</li> <li>5. <b>CAIXA Virtual Workshop</b><sup>40</sup>: June 22, 2020</li> </ol>

<sup>36</sup> <https://ibidaas.eu/blog/I-BiDaaS-Info-Day-Workshop>

<sup>37</sup> <https://www.ibidaas.eu/blog/I-BiDaaS-CRF-Hackathon/>

<sup>38</sup> <https://ibidaas.eu/events/I-BiDaaS-at-the-BDV-PPP-Summit-2019/>

<sup>39</sup> <https://ibidaas.eu/events/I-BiDaaS-at-the-European-Big-Data-Value-Forum/>

<sup>40</sup> <https://ibidaas.eu/blog/I-BiDaaS-Application-to-the-Financial-Sector-Workshop/>

	<p>6. <b>BDV PPP Virtual Summit 2020<sup>41</sup></b> - Big Data Pilot Demo Days series of webinars (Collaboration between I-BiDaaS and BigDataStack) May 21 – July 16, 2020</p> <p>a. <b>I-BiDaaS Application to the Financial Sector, May 21, 2020<sup>42</sup></b></p> <p>b. <b>I-BiDaaS Application to the Telecommunication Sector, June 25, 2020<sup>43</sup></b></p> <p>c. <b>I-BiDaaS Application to the Manufacturing Sector, July 9, 2020<sup>44</sup></b></p> <p>7. <b>Telefonica Research Online Hackathon<sup>45</sup></b>: October 23-25, 2020.</p> <p>8. <b>European Big Data Value Forum 2020<sup>46</sup></b>: November 3, 2020 (Collaboration between I-BiDaaS, BigDataStack and Track&amp;Know)</p> <p>9. <b>I-BiDaaS Final Event<sup>47</sup></b>: December 21, 2020.</p>
<b>KPI-DC-9</b>	<p>Five (5) collaborations with H2020 projects have been initiated:</p> <ol style="list-style-type: none"> <li>1. <b>TOREADOR</b> - Trustworthy model-aware Analytics Data platform (GA #688787)<sup>48</sup></li> <li>2. <b>DataBench</b> - Evidence-Based Big Data Benchmarking to Improve Business Performance (GA #780966)<sup>49</sup></li> <li>3. <b>BigDataStack</b> – Holistic Stack for BigData Applications and Operations (GA #779747)<sup>50</sup></li> <li>4. <b>Track&amp;Know</b> – Big Data for Mobility Tracking Knowledge Extraction in Urban Areas (GA # 780754)<sup>51</sup></li> <li>5. <b>Infinitech</b> - Tailored IoT &amp; BigData Sandboxes and Testbeds for Smart, Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem (GA # 856632)<sup>52</sup></li> </ol>
<b>KPI-DC-10</b>	<p>I-BiDaaS is actively participating in most (if not all) BDVA and ENISA<sup>53</sup> activities, I-BiDaaS Coordinator Dr. Sotiris Ioannidis is a PSG member of ENISA. Moreover, I-BiDaaS is supporting EU's purpose to become a global leader in accelerating digital transformation and contributing to Europe's determination in transitioning to circular industries, Dr. Giorgos Demetriou is a contributor of Veltha<sup>54</sup>, member of the Screen Policy Lab<sup>55</sup>, and member of the Circular Economy Expert Group<sup>56</sup> from the European Commission. Finally, UNSPMF is a member of the European Consortium for Mathematics in Industry (ECMI). Several activities related with ECMI included I-BiDaaS participation, e.g., the BIGMATH project advanced course 4.</p>

***I-BiDaaS contribution towards each of the expected impacts mentioned in the work programme:*** The motivation behind the targeted call was to develop technologies that would increase the efficiency of all EU companies and organisations that need to manage vast and complex amounts of data and in particular the competitiveness of EU enterprises. The emphasis was on rigorously measured increases in performance in data processing at a very large scale. More specifically, the I-BiDaaS contributions are summarised below:

- *Powerful (Big) Data processing tools and methods that demonstrate their applicability in real-world settings, including the data experimentation/integration (ICT-14) and Large Scale Pilot (ICT-15) projects.* I-BiDaaS developed and delivered a set of

<sup>41</sup> <https://www.ibidaas.eu/blog/Reaping-the-benefits-of-Big-Data-developments:-Big-Data-Pilot-Demo-Days/>

<sup>42</sup> <http://www.ibidaas.eu/events/Big-Data-Pilot-Demo-Days%3A-I-BiDaaS-Application-to-the-Financial-Sector>

<sup>43</sup> <http://www.ibidaas.eu/events/I-BiDaaS-Application-to-the-Telecommunication-Sector-Webinar>

<sup>44</sup> <http://www.ibidaas.eu/events/I-BiDaaS-Application-to-the-Manufacturing-Sector-Webinar>

<sup>45</sup> <https://www.ibidaas.eu/blog/Telefonica-Research-Online-Hackathon/>

<sup>46</sup> <https://www.ibidaas.eu/blog/European-Big-Data-Research-for-Industry-Report-online-now/>

<sup>47</sup> <https://youtu.be/xmIUGSxfnLQ>

<sup>48</sup> <http://www.toreador-project.eu/>

<sup>49</sup> <https://www.databench.eu>

<sup>50</sup> <https://bigdatastack.eu>

<sup>51</sup> <https://trackandknowproject.eu>

<sup>52</sup> <https://www.infinitech-h2020.eu>

<sup>53</sup> <https://www.enisa.europa.eu/>

<sup>54</sup> <https://www.veltha.eu/>

<sup>55</sup> <https://screen-policy-lab.mn.co/>

<sup>56</sup> <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3517>



complementary data processing tools, applicable to both batch and streaming data. The impact measurement indicators and targets that have been achieved are related to KPI-IE-2, KPI-IE-3 and KPI-IE-4 that have been described in Table 25 and Table 26 above. I-BiDaaS delivered several standalone tools (TDF, AVT, Hecuba, Qbeast, ML pool of algorithms, etc.). 8 real-life experiments have been performed that demonstrate the applicability of the I-BiDaaS solution to address 8 real problems in the targeted sectors and two generic experiments have also been defined aiming to demonstrate how the I-BiDaaS solution can be applied generally. The I-BiDaaS Consortium organised nine (9) special events (Workshops, Info Days, Hackathons, Webinars) and participated to even more events (BDV PPP Summit 2019, EBDVF 2019, BDV PPP Summit 2020, EBDVF2020) aiming to establish collaborations for further applicability. Different SMEs have been engaged that closely follow all the I-BiDaaS updates.

- *Demonstrated, significant increase of speed of data throughput and access, as measured against relevant, industry-validated benchmarks.* The impact measurement indicators and targets that have been achieved are related to KPI-RI-2 and KPI-RI-3. The use case providers report the achievement of significant improvements within I-BiDaaS. For CRF, the time needed to analyse data and make decisions was 1 month for both use cases. With I-BiDaaS, the time that is now needed has been reduced to 1 day. For TID, with I-BiDaaS, they achieved an increase of 200% in the number of detected low customer satisfaction calls by human agents. For CAIXA, before I-BiDaaS, the time to provide data to a new data analytics provider was 2 weeks and now they only need one day. It is also worth to mention that some numbers achieved by our data providers exceed by far the median business KPIs in the relevant sectors as identified by DataBench (see section 4). Moreover, the I-BiDaaS data providers (thanks to IBM's TDF) created new synthetic datasets that were not available before and also created new anonymised datasets. These datasets were 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. Especially for CAIXA, they managed to share tokenised data that made Big Data experimentation more agile, as experimenting in house may be time and resource consuming due to internal bank processes
- *Substantial increase in the definition and uptake of standards fostering data sharing, exchange and interoperability.* The impact measurement indicators and targets are related to KPI-RI-4, KPI-RI-5, KPI-DC-6 and KPI-DC-7. I-BiDaaS is contributing to a standard via BDVA. Moreover, In December 2018, a new standard, Recommendation ITU-T Y.3519, "Cloud computing – Functional architecture of big data as a service" was released. This standard seems to fully cover the I-BiDaaS functional architecture. Finally, KPI-DC-7 presents the lists with the standards for the use cases and the components that I-BiDaaS uptakes. A detailed report of these KPIs can be found in Table 25 and Table 26 and also below, where how I-BiDaaS contributes to standards and initiatives is described.

*Impact in research community and contribution to innovation capacity:* The I-BiDaaS project has achieved a significant impact on the research community. During the course of the project, I-BiDaaS researchers published 3 journal publications (33.3% gold-open access and 66.7% green open-access), 18 conference papers and 4 poster publications. We achieved 50% of these publications to have an impact factor or ERA classification. I-BiDaaS partners attended 24 conferences, 9 workshops and 32 other events. I-BiDaaS use case providers released to the research community 9 different big datasets (CAIXA: 3 real tokenised & 1 synthetic, TID: 2

synthetic, CRF: 2 real anonymised & 1 synthetic). In order to maximise the impact on the research community, all I-BiDaaS publications and datasets are available in Zenodo<sup>57</sup>.

One of the main goals within I-BiDaaS was to maximise the impact and deliver the innovations as close to the market as possible. The path from technology readiness to commercial exploitation, however, is a long, non-linear process. The technology readiness level (TRL) model is a systematic metric/measurement system that is used to assess the maturity of a technology when compared with various types of technology. The TRL maturity model comprises a scale of nine levels for describing the maturity. Level 1 and level 2 are related to knowledge development in academic levels. Level 3 to level 7 corresponds to technology developments derived from the collaboration between academia and industry. Level 7 to level 9 include industrial business development and are the ones closer to future development in the market.

Table 27 summarises the modules that have been integrated into the final version of the I-BiDaaS solution, the assessed TRL of each module, along with a comprehensive explanation highlighting the overall impact that has been achieved during the project. I-BiDaaS tools managed to rise to higher levels in the TRLs scale, bringing them closer to the market.

**Table 27: I-BiDaaS Tools & Technologies – Achieved Technology Readiness Level**

Tool	Partner	Initial TRL	Reached TRL	Impact
<b>GPU accelerator technology</b>	FORTH	3	5	The GPU-accelerated pattern matching engine has been offered as an API for data analytics, offloading computationally intensive tasks to GPUs for acceleration. The current version provides different types of analytics (such as sentiment score, word frequencies, and most frequent words) for different time intervals (e.g., last minute, last hour, and last day).
<b>COMPSs (Programming Model and Runtime)</b>	BSC	9	9	COMPSs was already TRL9 before the project start. During the project, COMPSs has gained the experience of working with real applications from the financial sector. Besides, the integration of COMPSs with Hecuba has been enhanced, supporting new data types and caching mechanisms.
<b>Hecuba</b>	BSC	5	6	Although Hecuba had been validated before in scientific environments, I-BiDaaS has been the first time where Hecuba has been used for industrial use cases, demonstrating thus its value. The main developments achieved have been related to an enhanced integration with the dislib library, the exploitation of locality when using COMPSs, and a set of new functionalities (i.e., filtering capacities, dynamic task granularity) as well as performance improvements and bug fixes.
<b>Qbeast</b>	BSC	5	6	I-BiDaaS has been a big push to Qbeast due to the fact that the framework was in the precise moment of being transformed from a research tool to a commercial product. This means that the I-BiDaaS use cases have enormously helped in this process, providing real-world scenarios that have helped shaping Qbeast. More precisely, interactive analysis has been implemented through a new architecture that integrates COMPSs, dislib and Hecuba with the Jupyter ecosystem (Voilà and Widgets), and javascript frameworks. This enables users to break the classical linear data analysis workflow to a new interactive loop where feedback from users can

<sup>57</sup> <https://zenodo.org/communities/i-bidaas/>

				trigger extra analyses while updating their models and algorithms.
<b>Test Data Fabrication</b>	IBM	6	7	TDF tool was used during the I-BiDaaS project for the fabrication of synthetic data for all of the use cases. The tool has been significantly enhanced to handle the use cases requirements, e.g., support for data streaming, automation of fabrication rules definitions, parallel CSP solver and a new GUI with improved modelling of data fabrication rules and improved usability.
<b>Apama Streaming Analytics Platform</b>	SAG	6	7	It has enabled the streaming use cases, either directly or by passing information back and forth between Terracotta DB and Apama and between Python and Apama. We can now react within a few seconds to anomalies and possible frauds. On the other hand, the handling of Apama when interfacing with Terracotta DB and Python is much clearer now. Related information has been forwarded to the R&D department of SAG so that the tool can be further improved.
<b>Universal Messaging</b>	SAG	7	8	We enabled the sending of MQTT messages to UM from IBM Data Fabrication Platform and from the Hecuba Batch Tools. So all data between the I-BiDaaS components is exchanged via UM. On the other hand, we have a clearer understanding now how messages can be kept for a defined duration so that consumers will see older ones within a given time frame when connecting.
<b>Pool of ML algorithms based on structured (non)convex optimization</b>	UNSPMF	3	5	Several developed algorithms have been validated on real industrially-relevant data provided by I-BiDaaS data providers, and the algorithms have been validated by the data providers. For example, the algorithms developed by UNSPMF have been validated for the following use cases: Accurate location prediction with high traffic and visibility (TID); Optimization of Placement of Telecommunication Equipment (TID); Analysis of relationships through IP addresses (CAIXA); Maintenance and Monitoring of production assets (CRF); Production Process of Aluminium die-casting (CRF).
<b>Orchestrator</b>	ITML	-	5	The orchestrator is a middleware between the AVT and the rest of the components. It exposes a REST API that is used by AVT to create new projects and experiments in all I-BiDaaS modes (Expert, Self-Service and Co-Develop mode).
<b>Advanced visualization and monitoring</b>	AEGIS	4	6	The AEGIS Advanced Visualization Toolkit (AVT) consists of a set of interactive visualisation tools developed to allow for a more straightforward exploratory analysis. The selection of the tools and, subsequently, the definition of the interactions among them depend on the domain and on the end-users' requirements. AVT development has started in the context of I-BiDaaS and has been developed along the project to successfully support use cases in the domains of finance, manufacturing and telecommunications. AVT has reached TRL level 6 and via I-BiDaaS, it has managed to be listed under the EU innovation Radar <sup>58</sup> and the BDV Marketplace <sup>59</sup> . Moreover, an internal dedicated team has been formed to move the tool at a TRL8 level, also taking into account progress performed in the context of other EU-funded projects that AEGIS is currently participating. Finally, a license scheme together

<sup>58</sup> <https://www.innoradar.eu/innovation/35294>

<sup>59</sup> <https://marketplace.big-data-value.eu/content/aegis-advanced-visualization-toolkit-avt>

				with a solid process for production deployment (including deployment paths for clients in one of the I-BiDaaS domains but also cases in entirely different domains) are being currently established by AEGIS in view of the commercialization of AVT.
<b>Resource management and orchestration module – Adaptation Engine Submodule</b>	ATOS	2	5	The Adaptation Engine provides a configurations management system capable to automate the lifecycle of the cloud-based deployments, it includes a catalogue of predefined automation recipes, configuration files and pre-packaged containers and VM templates. Within I-BiDaaS, it has been deployed on D-Alix datacentre located on the grounds of ITER institute of Science and Technology.
<b>Resource management and orchestration module – Cloudify Cloud Orchestrator Submodule</b>	ATOS	9	9	Cloudify is delivering real-world cloud native transformation across industries such as ATOS company using it to accelerate application development and DevOps practices. <i>Workflows</i> in Cloudify execute operations that are declared in TOSCA types. As such, <i>workflows</i> search the blueprint node graphs for nodes that refer to interfaces of interest to that workflow. Together with a set of blueprints and VM templates, it has been used in I-BiDaaS to orchestrate private and public cloud providers while enables the deployments across other infrastructure technologies such as Docker containers.
<b>Resource management and orchestration module – Private Cloud based on Openstack Submodule</b>	ATOS	9	9	OpenStack is a set of software tools for building and managing cloud computing platforms for public and private clouds. Backed by some of the biggest companies in software development and hosting, as well as thousands of individual community members. Within I-BiDaaS, it is installed on top of the bare metal servers, three different distributions have been used during the project life-time; DevStack, Kolla, for cloud deployments and MicroStack to proof edge/IoT deployments.

Moreover, an important achievement for the I-BiDaaS project is 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. To promote all the innovation results developed within the course of the project to a broader audience, as well as to help carry out matchmaking for potential future users of the I-BiDaaS platform and its components, all the innovations were also submitted to the BDVA Innovation Marketplace and after the review process, they were all accepted. The full list of the accepted Innovations is depicted in Table 28:

**Table 28: List of I-BiDaaS Excellent Innovations as accepted from EU Innovation Radar<sup>60</sup>**

Innovation	Key Innovators	Market Maturity	Innovation Topic
<b>Multidimensional Storage with Efficient Sampling (MuSES)<sup>61,62</sup></b>	BSC	<b>Market Ready</b>	<b>Deep Tech</b>

<sup>60</sup> <https://www.innoradar.eu/resultbykeyword/I-BiDaaS>

<sup>61</sup> <https://www.innoradar.eu/innovation/35296>

<sup>62</sup> <https://marketplace.big-data-value.eu/content/multidimensional-storage-efficient-sampling-muses>

<b>Advanced Visualization Toolkit (AVT) supporting scalable data visualisation<sup>6364</sup></b>	AEGIS	<b>Business Ready</b>	<b>Deep Tech</b>
<b>ADMM Machine Learning Algorithms<sup>6566</sup></b>	BSC UNSPMF	<b>Exploring</b>	<b>Deep Tech</b>
<b>Parallelization of constraint satisfaction problems<sup>6768</sup></b>	IBM	<b>Exploring</b>	<b>Deep Tech</b>
<b>Specification of an end-to-end Big Data as-a-self-service platform<sup>6970</sup></b>	UNIMAN ATOS UNSPMF	<b>Exploring</b>	<b>Deep Tech</b>

An ambitious goal set in the early stages was to achieve promising high TRL demonstrations and a feasible solution to meet the objectives of the project. To this end, the I-BiDaaS partners managed to develop a solution applied to the three different sectors (banking, telecom, manufacturing) and to the eight (8) industry-led experiments that reached TRL6 (technology demonstrated in relevant environment) and TRL7 (system prototype demonstration in operational environment). In Table 29, we report for each use case developed within I-BiDaaS the achieved TRL with a concise explanation.

**Table 29: I-BiDaaS Use Cases – Achieved Technology Readiness Level**

No.	Use Case	TRL	Explanation
1	Accurate location prediction with high traffic and visibility	6	The use case was tested in a relevant industrial environment, where the data were sourced from a major European telecommunications operator and fed into the relevant I-BiDaaS platform modules in an in-house installation. More specifically, we tested several months of anonymised mobility data in our predictive modelling setup and replicated previously reported results on the models' accuracy. The demonstration of the said technology was held for both TID and the EU telecom operator expert personnel, and their feedback was received.
2	Optimization of placement of telecommunication equipment	6	The use case was tested in a relevant industrial environment, where the data were sourced from a major European telecommunications operator and fed into the relevant I-BiDaaS platform modules in an in-house installation. More specifically, we tested several months of antenna network performance data in our predictive modelling setup and replicated previously reported results on the models' hotspot prediction accuracy. The demonstration of the said technology was held for both TID and the EU telecom operator expert personnel, and their feedback was received.
3	Quality of Service in Call Centers	6	The use case was tested in a relevant industrial environment, where the data were sourced from various Call Centers of Telefónica and fed into the relevant I-BiDaaS platform modules in an in-house installation. More specifically, we tested more than 1M of call transcripts in our predictive modelling setup and replicated previously reported results on the models' CSI prediction accuracy. The demonstration of the said technology was held for several relevant BI in TID, and the feedback from expert personnel was received.

<sup>63</sup> <https://www.innoradar.eu/innovation/35294>

<sup>64</sup> <https://marketplace.big-data-value.eu/content/aegis-advanced-visualization-toolkit-avt>

<sup>65</sup> <https://www.innoradar.eu/innovation/35298>

<sup>66</sup> <https://marketplace.big-data-value.eu/content/admm-machine-learning-algorithms>

<sup>67</sup> <https://www.innoradar.eu/innovation/35295>

<sup>68</sup> <https://marketplace.big-data-value.eu/content/parallelization-constraint-satisfaction-problems-csp-solver>

<sup>69</sup> <https://www.innoradar.eu/innovation/35293>

<sup>70</sup> <https://marketplace.big-data-value.eu/content/specification-end-end-big-data-self-service-platform>

4	Enhance control of customers to online banking access	7	Use case 4 and use case 5 achieve TRL7 thanks to the usage of real tokenized data. Although the process of analysis of this data was not completely integrated into the internal systems and procedures of CaixaBank, those use cases used were tested in a controlled by real-world scenario with real data, and were analysed by the digital security and Security Operations Center personnel in order to identify potential frauds or bad practices (as they do with any other potentially fraudulent transaction coming from other sources).
5	Advanced analysis of bank transfer payment in financial terminal	7	
6	Analysis of relationships through IP addresses	6	This use case was tested in a laboratory environment with synthetic and real tokenized data. The use case was mainly planned as a technology demonstrator and it was actually used as the I-BiDaaS MVP demonstrator at the beginning. The use case and the data are realistic, although it was not tested in a production environment and the results were not integrated in the customer social graph database.
7	Maintenance and monitoring of production assets	6	The use case has a TRL 6 because it has been tested in a relevant industrial environment (data coming from the industrial plant, use in the batch analytics, testing in Campus Melfi). The I-BiDaaS solution has been tested through a case study in which Big Data, gathered from sensors mounted on several machines along the production line of vehicles, have been analysed. The analysis was based on the development of an anomaly detection model for the detection of sensors' anomalies. With the results, a foundational database has been built and it allows to easily visualise the history of the trend of anomalies over time. All of this has been tested in a relevant industrial environment by CRF and allowed CRF experts to test the efficiency of the results for the maintenance planning to avoid failures before they happen.
8	Production process of aluminium die-casting	6	The use case has a TRL 6 because it has been tested in a relevant industrial environment (data coming from the industrial plant, use in the stream analytics, testing in Campus Melfi). The I-BiDaaS solution has been tested through a case study in which Big Data, gathered from the production process of the engine blocks, have been analysed. The analysis was based on the development of a Machine Learning algorithm for the detection of the quality level of the engine block. CRF experts tested in a relevant industrial environment the I-BiDaaS solution through the Advanced Visualisation Tools developed within the project. Specifically, it is possible to timely visualise the classification levels of the quality of the engine block in the preliminary stage of the process and avoid further processing and scraps thanks to the possibility to act before the next production steps.
9	Experiment for end-to-end I-BiDaaS platform in self-service mode	6	For the self-service mode, the integrated I-BiDaaS solution has been demonstrated in a relevant environment. For example, it has been demonstrated as an online platform and evaluated by external personnel, including CAIXA employees (CAIXA workshop), Python developers of FORTH, UNSPMF, ATOS, CRF, AEGIS, ITML, etc.
10	Experiment for end-to-end I-BiDaaS platform in expert mode	6	For the expert mode, the I-BiDaaS integrated solution has been demonstrated in the CAIXA environment, on real CAIXA tokenized data. Moreover, some CAIXA use cases have been implemented in the expert mode.

*Impact in Data Market and the Big Data Economy:* I-BiDaaS is a Big Data as a Self-Service Solution that provides a significant boost to the financial, 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. Below, the main benefits for each data provider are presented.

The main benefits obtained by CAIXA due to its participation in I-BiDaaS are highlighted in Table 30.

**Table 30: Summary of the impact of the CAIXA use cases studied in I-BiDaaS**

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

Through I-BiDaaS, CAIXA was able to speed up the implementation of Big Data analytics applications, test algorithms outside CAIXA premises and test new tools and algorithms without data privacy concerns by exploring and validating the usage of synthetic data and tokenised data in three different use cases, improving the efficiency in time and cost by means of skipping some data access procedures and being able to use new tools and algorithms in a much more agile way. User requirements regarding the availability of ‘Intermediate and Non-IT users’ to analyse and process the data of the use cases were also validated through several internal and external workshops in which the attendants from several departments of CAIXA and other external entities (data scientists, business consultants, IT and Big Data managers) provided very positive feedback about the platform usability. Moreover, use cases “Advanced Analysis of bank transfer payment in financial terminal” and “Enhance control of customers to online banking”, as mentioned previously, were also validated by the corresponding business processes employees, being able to extract the results by themselves.

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

The main benefits obtained by CRF are reported in Table 31:

**Table 31: Summary of the impact of the CRF use cases studied in I-BiDaaS**

Benefits	KPIs
To enhance production times, to reduce costs and, consequently, to obtain results that satisfy manufacturers’ requests in terms of product quality, machine performance and timing	Increase of 3-7 % of quality control level related to good products and decrease of 1-4% and 0.5-2 % of two quality control levels related to defective products
To improve the efficiency of manufacturing plants, by getting the best performance from the machinery to reduce production losses and achieve greater competitiveness of the company	Increase of 1-1.5 % in current Overall Equipment Effectiveness (OEE) and decrease of 20-30 % in maintenance costs
To reduce time to produce decisions	From one month to few hours, a turn of job or 1 day
To achieve accuracy of new models with respect to internal CRF models in use	Increase of 6 % for the first use case and 20% for the second one

The I-BiDaaS solution gave to CRF the insight and tools to develop a methodology for the implementation of Big Data analytics in the manufacturing sector, in those areas in which data are generated to set up and control equipment and processes. First of all, data, gathered from several sources and several levels, have to be ingested to understand how they have to be managed, depending on whether we want to act on the quality of the production process or on the maintenance of the equipment. Furthermore, the complexity of the industrial processes requires that different departments and different figures are involved to understand and extract the value, often hidden in raw data. It is important to break data silos through the collaboration of those who have the relevant experience (manufacturers), those who are involved in data collection, data security, manual analysis and operational flows (intermediate users) and those who are employed at different levels in production processes (operators). After identifying data useful for the analysis, another important step is the anonymisation of them for security reason. In our case anonymisation required more time than expected, so we understood that it can be useful to fabricate realistic synthetic dataset for the early stage of the implementation. Then, the data cleaning has been the next step that allowed to identify incomplete, inaccurate and irrelevant parts of the generated dataset and to understand which types of data can be useful for the analysis. When the datasets were ready, we selected which Big Data technologies are most suitable for the specific identified business requirements. Batch and stream analytics cover all aspects, which may occur in real-world environments, including 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). Finally, the advanced visualisation tools, developed within the I-BiDaaS project, provided the insights, value, and operational knowledge extracted from available data giving us the possibility to test the I-BiDaaS solution for our use cases.

The main benefits obtained by TID are reported in Table 32. Business Units inspect <1% of total calls per year. The current solutions are both costly and time-consuming. However, through a number of simple, intuitive, and effective visualizations and dashboards for the end-users, the I-BiDaaS platform can facilitate a more effective operation of the call centres (improvement of QoS, automation, etc.) and support mining capabilities for monitoring customer satisfaction. As shown in Table 32, the proposed service has the capacity to increase the number of low customer satisfaction audio calls detected by human agents to 7,000 (a notable 200% increase) by pre-processing/filtering the audio calls and flagging those cases of interest. This solution goes a long way into making the CC more efficient, effective and, most importantly, cutting down our operational costs.

**Table 32: Summary of the impact of the TID use cases studied in I-BiDaaS**

Benefits	KPIs
~ <b>11,520</b> calls processed (per year) by a single human agent.	~ <b>3.5B</b> calls processed per year in a single GPU; max. real time throughput: 40K transcripts/second.
The manual process allows to identify about <b>2,300</b> low customer satisfaction audio calls.	Increase the number of detected low customer satisfaction calls by human agent to <b>7,000 (200% increase)</b> by ranking the calls based on sentiment.

*Contribution to standards and international initiatives:* The impact measurement indicators and targets are related to KPI-RI-4, KPI-RI-5, KPI-DC-6, KPI-DC-7 and KPI-DC-10. Four (4) members of the Consortium are founding members of the Big Data Value Association (BDVA). I-BiDaaS is fully aligned with all PPP BDVA activities like the European Big Data Value Strategic Research and Innovation Agenda (BDVA SRIA). Strong collaborations were pursued with other activities of the PPP that were directly related to the “data processing architectures”,



the “data analytics” and the “data visualisation and user interaction”. The I-BiDaaS solution can be contextualised within the BDV reference model defined in the BDV Strategic Research and Innovation Agenda (BDV SRIA) and contributes to the following technical priorities: Data protection; Data Processing Architectures; Data Analytics; and Data Visualisation and User Interaction. In December 2018, a new standard, Recommendation ITU-T Y.3519, “Cloud computing – Functional architecture of big data as a service” was released. This standard seems to fully cover the I-BiDaaS functional architecture. I-BiDaaS is participating and collaborating with DataBench project, which is designing performance benchmarking processes for Big Data. DataBench is expected to set the standards and benchmarks for the emerging Big Data ecosystem. Finally, in Annex II of this deliverable, several tables providing detailed representation of the standards used in the various use cases and components of I-BiDaaS are included.

The I-BiDaaS Consortium comprised complementary means and tools to ensure impact maximisation. Such means include the External Advisory Board, coordinated exploitation of project results, industrial strategic exploitation plans, knowledge providers’ exploitation plans and a sound communication and dissemination strategy.

The I-BiDaaS External Advisory Board significantly affected the overall progress of the project. The main task of the EAB was to provide external, independent analysis and recommendations on the project achievements and to bring additional competencies towards a full achievement of the I-BiDaaS objectives. The I-BiDaaS EAB includes Nuria de Lama Sanchez (European Programs Manager, Atos Research & Innovation - Member of Board of Directors, BDVA, Spain), George Vouros (Professor, Department of Digital Systems, University of Piraeus, Greece), Ilija Susa (Co-founder of Content Insights LLC, Serbia) and Jean-Marie Hurtiger (CEO of Desmond sas Automotive Consulting, President of RENAULT light Commercial Vehicles, and former CEO of Renault Samsung Motors (Korea), France).

Three meetings between the I-BiDaaS Consortium and the EAB members were organised. The first meeting was held online in October 2018. The main discussions and recommendations of the EAB members were about the data integration and ingestion regarding cross-domain flow from different stakeholders and the management of confidentiality and protection of proprietary information along with GDPR regulation compliance and data sharing. The second meeting with the EAB members was held on March 7th, 2019, at the premises of IBM, in Haifa, Israel. The feedback received during the meeting and I-BiDaaS work to address the comments/suggestions are summarised in Table 33. This meeting was crucial and it marked a turning point for the further development of the project.

**Table 33: EAB Comments and how I-BiDaaS works towards addressing them**

Name of the EAB member	Comment/Suggestion	How I-BiDaaS works towards addressing the comment/suggestion
Nuria de Lama Sanchez, George Vouros, Ilija Susa	Focus on one type of user (non-IT or IT user) for the next release of the I-BiDaaS solution (M18). Distinguish the different uses and users of the platform.	A clear distinction between the users and uses of the platform has been provided. We now make clear that there are three modes of operation of the platform: 1) <i>Expert mode</i> : allows the experts (developers) to upload their own data analytics code based on the available I-BiDaaS highly reusable templates. 2) <i>Self-service mode</i> : allows the users, having the relevant domain knowledge and some experience about data analysis (non-experts), to easily construct Big Data pipelines in a user-friendly way, selecting a pre-defined data analytics algorithm from the available list. 3) <i>Co-develop mode</i> : corresponds to an end-to-end solution for a given industry project developed by the I-BiDaaS team (the I-BiDaaS use cases).
Nuria de Lama Sanchez	Use BDVA Toolbox when it will be released to run I-BiDaaS benchmarks	I-BiDaaS and DataBench established an active collaboration. The I-BiDaaS data providers participated in a survey carried out by DataBench on Big Data use and impacts. I-BiDaaS was

		invited to participate in two webinars organised by DataBench to present the I-BiDaaS benchmarking approach, landscape and needs, and how the I-BiDaaS experimentation maps to the generic Big Data pipelines described in DataBench. DataBench provided to I-BiDaaS several recommendations on concrete benchmarks. I-BiDaaS contributed to the DataBench project's ReachOut campaign (see KPI-DC-6).
George Vouros	Measure throughput and latency of the I-BiDaaS platform in the next release.	Throughput and latency of the platform have been measured and reported in section 3.2 of this deliverable.
Nuria de Lama Sanchez, George Vouros, Jean-Marie Hurtiger, Ilija Susa	I-BiDaaS should guarantee the quality of the generated data. Suggestion: Benchmark the quality of the fabricated data by applying analytics on both real and fabricated data. Mitigation Plan: Metrics to compare the quality of the fabricated data with respect to the real data.	In D6.4 [4], an extensive evaluation of the quality of the fabricated data has been carried out. Data quality from the perspective of assessing algorithm scalability and specific and general utility 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'.
Nuria de Lama Sanchez	Create a more realistic business model canvas in the context of I-BiDaaS Commercialization.	The project's final year generated an enriched input to individual exploitation plans of the innovations identified by the EU radar and a sound business plan for the I-BiDaaS solution., maximizing exploitation opportunities for individual partners and sustainability of the tools beyond the project's lifespan. The Dynamic Business Model (DBM) adopted for the project enabled a rapid prototyping approach allowing the consortium to use lean and agile principles appropriate to achieve a faster time to market and seize emerging business opportunities for I-BiDaaS.
George Vouros	I-BiDaaS should specify the Innovations that the project will contribute to the market.	Innovations identified at M18 and submitted to the EU Innovation Radar after the 1 <sup>st</sup> project review. Five innovations of the I-BiDaaS project were evaluated as excellent innovations by the EU Innovation Radar. One innovation was deemed market-ready, one business ready and three exploring. The institutions that were main developers of the five innovations were identified as key innovators and the innovations have been published in the EU innovation radar platform <sup>71</sup> .
Nuria de Lama Sanchez	I-BiDaaS can contribute to the BDVA PPP newsletter.	A continuous communication channel with BDV PPP has been developed to further enhance our dissemination efforts. Starting from July 2019, I-BiDaaS blog articles are being published in the BDVA PPP newsletter (more information can be found in D7.7 [16]).
Nuria de Lama Sanchez	I-BiDaaS should make a clearer project statement by providing a 'selling story'.	We have worked significantly to sharpen the I-BiDaaS "selling-story". The key points of the project are the self-service features and flexibility (see above the three modes) of the platform, as well as the breaking silos effort of the project (Info Days, Hackathons, making available tokenized/anonymised real open data sets, etc.)
Nuria de Lama Sanchez, George Vouros	I-BiDaaS should focus more on Data – Data availability, Data Sharing and Open Data. Make data (real or fabricated) open as soon as possible as part of the Open Research Data Pilot.	I-BiDaaS data providers successfully shared 12 datasets, 8 of which are real anonymised/encrypted and 9 open access <sup>72</sup> . This change is considered important and drastically affected the overall impact of the project.
Nuria de Lama	I-BiDaaS should focus on Standardization activities and	The Consortium activities with respect to standardization are monitored through KPI-RI-4, KPI-RI-5, KPI-DC-6, KPI-DC-7

<sup>71</sup> <https://www.ibidaas.eu/blog/EU-Innovation-Radar-Identifies-IBiDaaS-Innovations-and-Key-Innovators/>

<sup>72</sup> <https://www.ibidaas.eu/tools/>

Sanchez, George Vourros	define the Standards that will uptake and influence.	and the status report is included in Table 25 and Table 26 of this document.
Jean-Marie Hurtiger	Create a proforma demo to demonstrate I-BiDaaS.	The Consortium prepared several end to end platform demos that are available in the I-BiDaaS YouTube channel <sup>73</sup> . Moreover, available in the I-BiDaaS YouTube channel are the BigDataPilotDemoDays Webinars and the I-BiDaaS Final Event, where the pilots demonstrated all the use cases in a step-by-step fashion.
Ilija Susa	I-BiDaaS should define the open-source tools and release them to receive contributions on the open source software.	One of the goals of the I-BiDaaS consortium is to contribute to the Open Source communities to provide benefit to the European community. Open source version of the I-BiDaaS platform, including 5 components, has been published and 4 tools are provided as open source (see KPI-RI-1). I-BiDaaS contributes to open-source tools by making several posts to the I-BiDaaS git knowledge repository <sup>74</sup> .

The final EAB meeting was held online on July 7<sup>th</sup>, 2020. Draft versions of all presentations, the deliverables and the feedback received during the 1<sup>st</sup> project review were made available to the EAB members beforehand, aiming to receive the most valuable feedback and consultation possible. The I-BiDaaS Consortium, through a series of presentations, had the opportunity to showcase the progress achieved since the last face-to-face meeting in Haifa with all the scientific & technical advancements. The main focus was mainly on the I-BiDaaS self-service solution and its application to the financial, telecommunication and manufacturing sectors via a step-by-step demonstration of 10 different use cases. In addition, the comments received during the 1st project review meeting and our actions to address them were extensively discussed. Access to the I-BiDaaS platform to conduct their own experiments was given to the EAB members. Finally, the commercial offering of the I-BiDaaS solution was presented. The EAB members expressed their satisfaction regarding the progress achieved and provided valuable inputs and feedback towards the successful accomplishment of the project vision and goals. The main focus of the discussions was related to the final datasets that have been shared, the open source components and the release of the open source version of the platform, how datasets can be uploaded, benchmarking, the role of TDF and the quality of the synthetic data, breaking of the data silos, the kind of decisions that the visualisation is triggering, and the business models that have been developed, how we plan to exploit them and what will be the steps to go to the market.

The other two means that were used utilised were related to the exploitation and dissemination activities of the project. The I-BiDaaS partners recognized that a well-coordinated exploitation of the project results is essential for a sustained influence and penetration of the project outcomes to the market place. Within the context of WP7, the I-BiDaaS exploitation team worked together with the partners and managed to achieve comprehensive and well-coordinated individual and joint exploitation plans. Especially during the third year of the project, I-BiDaaS managed to further advance its exploitation strategy since the project's outcomes reached the required maturity peak for exploitation purposes. The DBM adopted for the project enabled a rapid prototyping approach allowing the consortium to use lean and agile principles appropriate to achieve a faster time to market and seize emerging business opportunities for I-BiDaaS. The exploitation team did a reality check of the business models and finalized the design and proposition of a sound business plan, including financial modelling and the potential revenue analysis streams. The exploitation team also identified possible circular business models in the

<sup>73</sup> <https://www.youtube.com/channel/UCCBVaMmNbS1NPzXTvPQOkAA/videos>

<sup>74</sup> [https://github.com/ibidaas/knowledge\\_repository](https://github.com/ibidaas/knowledge_repository)

project use cases, which are considerably insightful for the different sectors that need a clear plan to boost a range of new skills, including big data analysis.

Finally, regarding the I-BiDaaS Communication and Dissemination strategy as a mean to maximise the impact, a concrete communication methodology consisting of four phases was carried out. The phases were a) Starting point b) Inception Phase c) Implementation Phase and d) Monitoring and Improvement. The dissemination and communication activities were implemented in line with the strategy aiming to raise awareness of I-BiDaaS activities and make this project successful. Significant achievements have been accomplished and all objectives set in the dissemination and communication plan created for the project have been met. This is reflected by the key achievements reported through this report and the level completion of the KPIs (KPI-DC-1 up to KPI-DC-10) outlined in Table 25 and Table 26 of this deliverable. Especially during the third year of the project, a powerful presence on the web and social media was achieved. Moreover, a series of strong collaborations with other projects and initiatives have been developed. The dissemination strategy, along with the activities and the achievements of the I-BiDaaS project, can be found in D7.3 [15], D7.5 [14] and D7.7 [16].

### **Consolidation Phase - Progress report**

According to the original time plan of the project, after the end of M30 and the completion of the Experimentation and Evaluation phase, the I-BiDaaS project should enter into the Consolidation phase. However, due to the COVID-19 pandemic, there was a delay of two months. Thus, the project entered into the Consolidation phase at M33 after the successful completion of the Experimentation and Evaluation phase where all functionalities developed during the previous period were implemented on 8 (eight) real-life industrial scenarios in the I-BiDaaS targeted domains of telecommunication, finance and manufacturing and two more generic scenarios. The output of the Experimentation and Evaluation phase fed the Consolidation phase, which aimed to ensure the impact and exploitability of the project results by improving the usability of the I-BiDaaS tools.

During the Consolidation phase, the main focus was on fine tuning the I-BiDaaS approach on completing its set of features and fixing any defects identified through the implementation of the previous phase's experiments. The final integrated version of the I-BiDaaS solution was released just before the end of the project that successfully demonstrates 10 different use cases. Significant efforts have been made by the I-BiDaaS technology providers to apply fixes on the tools based on the feedback received not only from the use case providers but also from external entities that were engaged to be the I-BiDaaS evaluators. In Section 5.2, a detailed description of the feedback received from external stakeholders is provided.

## **5.2. Feedback from external stakeholders**

In this section, we present the external stakeholders' feedback received during the project period M33-M36; for earlier results, the reader may refer to D6.1 [6], D6.3 [3] and D6.4 [4]. In particular, we discuss the following important events at which the external feedback was collected: Telefonica Research Hackathon, EBDVF2020, and the I-BiDaaS Final Event.

In this section, we present the external stakeholders' feedback received during the project period M33-M36; for earlier results, the reader may refer to D6.1 [6], D6.3 [3] and D6.4 [4]. In particular, we discuss the following important events at which the external feedback was collected: Telefonica Research Hackathon, EBDVF2020, and the I-BiDaaS Final Event.

**Telefonica Research Hackathon.** TID organised an online Hackathon event between the dates October 23rd to October 25th 2020, on the "Quality of Service in Call Centers", a high-value use case for any company that wants to maintain a close customer relationship customers. In this Hackathon challenge, motivated by the I-BiDaaS EU project, we proposed the analysis of Call Center transcripts and the corresponding voice acoustic features to predict CSI. Such

services may support the Call Center in screening phone calls automatically and identifying efficiently problematic cases. Hence, the main task was to provide an automatic solution for analysing the calls and predicting customer satisfaction.

For the CSI prediction, we considered the use of deep neural architectures to perform early feature fusion of both prosodic and textual information. Convolutional Neural Networks were trained on a combination of word embeddings and acoustic features. We addressed the task as a binary classification i.e. “low” and “high” satisfaction prediction categories. We, further, investigated whether fully anonymized transcripts can impact the performance. For the purposes of the hackathon, we made available an anonymized call-centre dataset.

The official track of the challenge was the following:

- CSI detection, as self-reported by the customer, on anonymize data from text-based and/or prosodic features

We also encourage the participants to submit their proposals to the following unofficial tracks:

- CSI detection on non-anonymised data from text-based and prosodic features
- CSI detection from only text
- CSI detection from only prosodic

Each team was provided with access to a dedicated AWS server (Ubuntu, instance type “g3s.xlarge”, Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz, 31 GB memory, SSD 90 GB, Tesla M60 GPU card) that had pre-installed updated versions of popular ML/DL libraries, along with a training dataset and examples of Python scripts in Jupyter notebook format for benchmarking their final submission.

<input type="checkbox"/>	Name	Instance ID	Instance state	Instance type	Status check	Alarm Status	Availability zone
<input type="checkbox"/>	TRH-Team-1	i-0b8712e216987d260	Running	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-2	i-0d6dee82dacc64c39	Running	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-3	i-01ad28bf170ee7c07	Stopping	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-4	i-05899afdad7b2c0a4	Stopping	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-5	i-03818d5c0b636f9f3	Stopping	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-6	i-03809c9d9e0537b05	Running	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-7	i-03552bc2dc5eae00d	Running	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Team-8	i-073758caa7eb43ee	Stopping	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b
<input type="checkbox"/>	TRH-Template	i-0b09e19f1501fa0eb	Running	g3s.xlarge	2/2 checks ...	No alarms +	eu-west-2b

**Figure 16. Screenshot of AWS servers used for the purposes of the Telefonica Research Hackathon**

The submitted solutions were evaluated on an unseen test set of approximately 9,000 instances, using the following scoring function, which is the geometric mean of two metrics:

$$Score = \sqrt{F_{\beta} \cdot NR_{score}}$$

- $F_{\beta}$ : The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall.
- $NR_{score}$ : The Normalised Runtime (NR) is the total number of milliseconds that the submitted system requires to perform classification in the final anonymised test set, averaged over 10 trials, and normalised by the corresponding baseline Runtime, as computed on the same test set.

After a careful assessment of the submitted solutions, we determined the winning team: “Team 7 – ElArbustoDeLaDecision” (Dennis Doerrich, Yaroslav Marchuk). Their model outperformed our baseline solution in terms of the F-beta score and achieved an impressive final score of 0.47164.

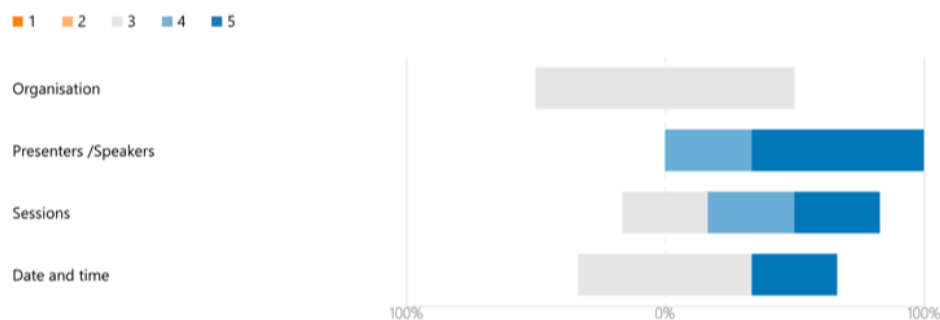


Team-7
Runtime (average of 10 rounds): 115.84413
Runtime score (compared to baseline): 0.50000
F-beta score (average of 10 rounds): 0.44488
Final score: 0.47164

Through this Hackathon event, TID achieved the following goals:

- Assisted in the synthetization of realistic data that mimic real datasets and facilitated the early exploration and development phases in I-BiDaaS
- Broke the inter- and intra-sectorial data-silos, and provided in-house access to real-life datasets
- Involved different business units and external companies for interfacing and exploring novel data analytic technologies
- Raised awareness about the challenges and research output produced by the I-BiDaaS project

In addition, we collected feedback from the participants at post-event and the questionnaire data was analysed by ENPC. When asked about their experience, the participants reported positive feedback, as shown in Figure 17. When asked about which elements of the event the participants liked the most, the participants reported that *“It was nice that you provided an AWS machine with data and code that we can use”* and *“The organizers were available and ready to help”*.



**Figure 17. Feedback to the question "Rate your event experience (being 5 the best rate experience)"**

With respect to the lessons learned, some participants commented *“Different approaches to an NLP problem”* and *“Insights on how a company manages and structures a code to solve a problem. How the industry tries to solve a use case”*. Among the biggest challenges, participants reported *“Getting a good enough score with simple models”*, *“Diving into the problem domain since our company does not focus on ML only”* and *“Engineering the response”*. Finally, when asked about the (positive) impact of the hackathon event on their business goals, a participant highlighted *“Totally, as already mentioned, it is good to see and understand how industrial companies adopt ML and emerging technologies.”*

In summary, this Hackathon addressed the challenge of developing speech technologies that transform audio calls into relevant information for the Call Center. By working synergistically on this use case, we were able to deliver technologies that can improve the number of audio calls processed per time unit and reduce significantly the manual effort allocated for this task. Also, the winning team was awarded with a free entrance to the Wayra pitch day.

**European Big Data Value Forum 2020 (EBDVF 2020).** The European Big Data Value Forum (EBDVF) is the flagship event of the European Big Data and Data-Driven AI Research and Innovation community organized by the Big Data Value Association (BDVA) and the European Commission (DG CNECT). The 2020 edition of the EBDVF took place between the 3rd and

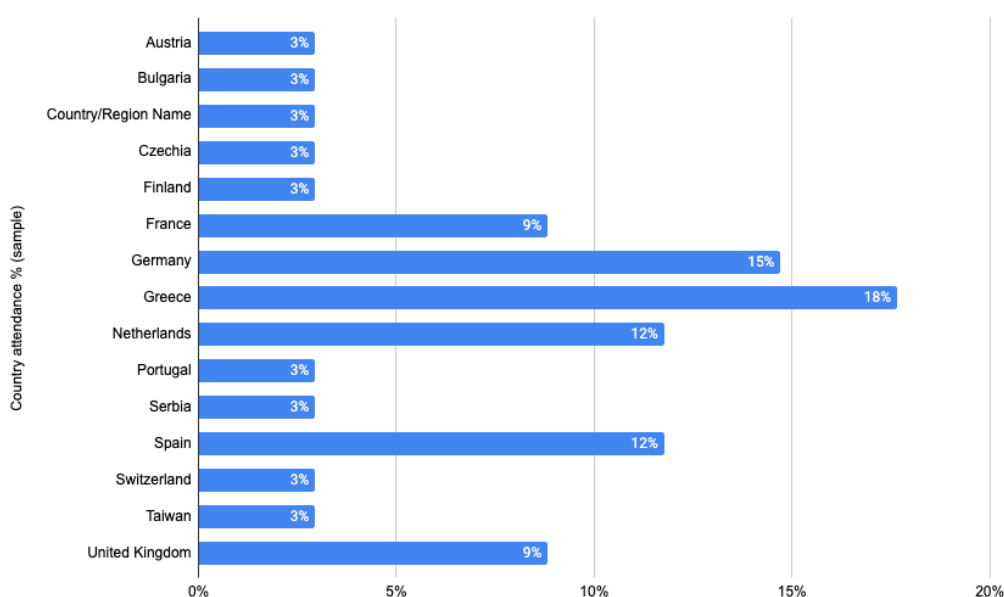
the 5th of November 2020. I-BiDaaS was sponsoring EBDVF2020 and managed to have a strong presence by co-organising a session and by actively participating to 3 (three) more sessions, having as a main goal to share the I-BiDaaS results and the lessons learned.

*1. Parallel session on European Big Data Research for Industry. 3 projects. 7 sectors. 9 applications. 41 software components. Now what? (I-BiDaaS sponsored event).*

On Tuesday, November 3<sup>rd</sup>, 2020, within the framework of the European Big Data Value Forum, Big Data EU H2020 research projects I-BiDaaS, BigDataStack and Track & Know hosted a joint session. The collaboration between the three EU projects was initiated at the beginning of 2020 when during the BDV PPP Summit 2020, they decided to join forces in a series of 9 online demonstrations of innovative Big Data Technologies in the pilot studies and their applicability to an ever wider scope contributing to Europe's digital future: the Big Data Pilot Demo Days. After 3 years of research and innovation, the projects joined forces again in the expert-led discussion on the impact and uptake of Big Data research results. The purpose of the discussion was to (i) identify shared barriers to adoption of Big Data research in different sectors, and mechanisms to overcome them, (ii) map the current and future impact and sustainability of their Big Data research, (iii) share best practices on the concrete business questions that have been answered in the project pilots.

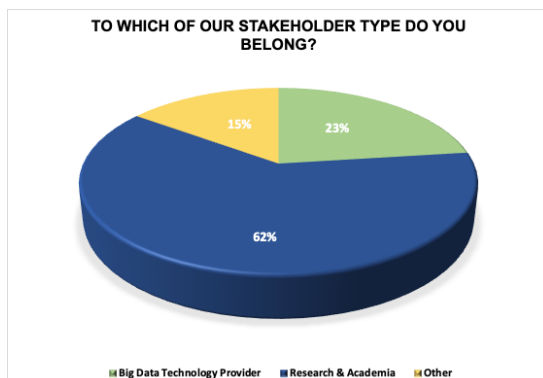
In total, 117 attendees from more than 14 countries attended the session (see Figure 18), a number that significantly increased the visibility of the projects.

EBDVF 2020 - Parallel session on European Big Data Research for Industry. 3 projects. 7 sectors. 9 applications. 41 software components. Now what?

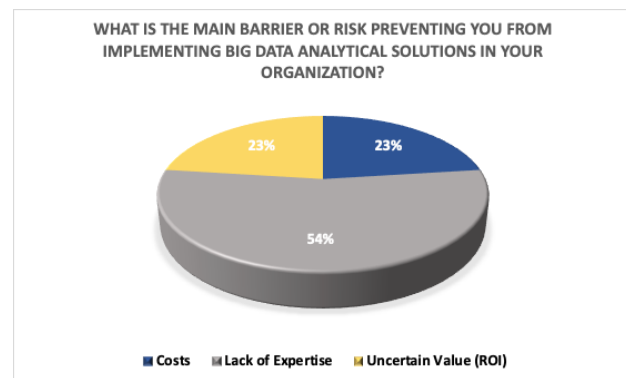


**Figure 18. EBDVF2020 - Geographical Spread**

Fostering engagement, the attendees were asked a few questions in order to understand their background and interests. From the responses received, the majority of the attendees were from Research and Academia (62%), followed by Big Data Technology providers (23%), as shown in Figure 19. An overwhelming 92% worked with Big Data and 85% were interested in Big Data technologies to improve their customer experiences, the main barrier to achieve this was considered the lack of skills (54%) as shown in Figure 20.



**Figure 19. To which of our stakeholder types do you belong?**



**Figure 20. What is the main barrier or risk preventing you from implementing Big Data Analytical solutions in your organisation?**

The result of this group discussion is an insightful online report with key findings with recommendations on how to facilitate the uptake of results and to make Big Data research more sustainable and future proof<sup>75</sup>.

The group discussion focused on 3 key questions:

- How are we contributing to the European Big Data Ecosystem?
- How can we apply Big Data in businesses? What barriers did you encounter with the industries in your project?
- Now what?

The expert panel was composed of three (3) invited speakers, one from each project:

- I-BiDaaS: Alon Rozen - Professor of Innovation, Dean, Ecole des Ponts Business School, I-BiDaaS Exploitation manager,
- Track & Know: Toni Staykova - Co-Founder and Vice President, UKeMED,
- BigDataStack: Richard McCreadie - Lecturer in Information Retrieval and Data Systems, University of Glasgow.

The I-BiDaaS Exploitation Manager, Prof. Alon Rozen, highlighted how I-BiDaaS contributes to the European Big Data Ecosystem as an actor and enabler of (i) an EU big data ecosystem by contributing itself to the ecosystem as it is both an enabler, a component, and a catalyst while looking for a solution to enlarge the ecosystem, (ii) EU big data innovations: 11 software components, five of which are open source, 10 European use cases across three different sectors, five innovations that were deemed mature by the EU Innovation Radar, lessons learned on data sharing and data availability and (iii) democratization of Big Data solutions to EU SMEs: two innovations were deemed business ready by the EU Innovation Radar, several other are moving in that direction too.

Regarding the ‘what’s next?’ aspect, Prof. Rozen described the key outcomes in different levels: (i) project-level: the different modes of I-BiDaaS solution and the Cookbook/User Guide help on further adoption by any organisation, (ii) consortium level: the consortium will continue to live and work together as an agile ecosystem, take the innovations to the market, help them mature, climb the experience and learning curve by training and upskilling in-house talent and (iii) geopolitics: I-BiDaaS consortium is building a bottom-up emergent big data ecosystem, a fractal of the bigger EU ecosystem.

<sup>75</sup> <https://doi.org/10.5281/zenodo.4326876>



## 2. *Evaluation schemes for Big data and AI Performance of high Business impact (DataBench)*

I-BiDaaS and DataBench projects have established a productive collaboration. As a result, DataBench project invited I-BiDaaS to participate to the “Evaluation schemes for Big data and AI Performance of high Business impact” session organised by DataBench, which took place on November 4th, 2020. The I-BiDaaS perspective on Big Data and AI architectural pipelines and benchmarks was presented. In more detail, a thorough presentation of the I-BiDaaS project and its experimental workflow using the I-BiDaaS application in the banking sector as a representative example was given (See section 3.4).

## 3. *Big Data Value Best Success Stories 2020*

BDVe project and BDVA have organized this year the second edition of the Best Success Story (BSS) contest. The contest was originally organised for the BDV PPP Summit 2020. Even though the event due to the COVID-19 pandemic was cancelled, the BSS contest took place as originally planned. Motivated by the valuable feedback received during the 1st project review, CAIXA decided to participate in order to promote and give visibility to the results produced. The CaixaBank’s Success Story in I-BiDaaS titled “Towards open and agile Big Data Analytics in Financial Sector”<sup>76</sup> was submitted. The main focus of the story included lessons learned on data sharing and data availability and how CAIXA completely changed its approach from a non-sharing real data at all position to looking for the best way possible to share real data and perform big data analytics outside its facilities. 21 stories in total were submitted. Although CAIXA’s BSS was not the final winner, it was selected as a runner up<sup>77</sup>. The CAIXA’s story was presented during the EBDVF 2020 on a dedicated slot for the Best Success Story Award on November 5<sup>th</sup>, 2020.

## 4. *Parallel session on Data Driven AI for Financial Services (Infinittech)*

CAIXA was one of the session invited speakers of the Parallel session on Data Driven AI for Financial Services which was organised by the EU H2020 Infinittech project on November 5<sup>th</sup>, 2020. During the session, the I-BiDaaS approach as an innovative solution for real-life scenarios from the Banking Sector was presented in order to provide insights on “How to exploit your big data and overcome the challenges and constraints of a highly regulated sector?”. The I-BiDaaS architecture, CAIXA’s use cases and the results obtained in the project were presented.

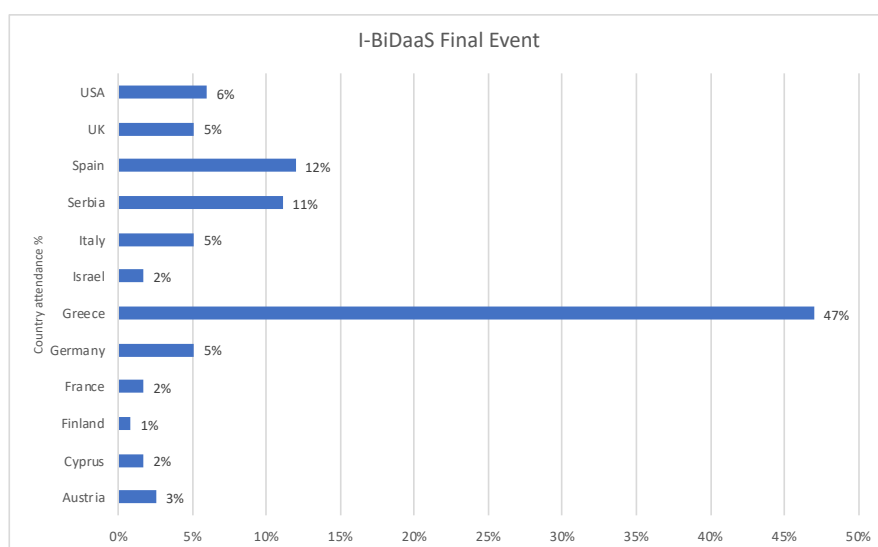
Moreover, during EBDVF2020, the projects sponsoring the event had several communication, interaction and engagement features. I-BiDaaS had its own virtual booth, which reached a significant number of views.

**I-BiDaaS Final Event.** After three years of research and innovation, the I-BiDaaS project partners organised a free online event that was held on the 21<sup>st</sup> of December 2020 in order to share the main results achieved. As this exciting journey is approaching the end and aiming to ensure further adoption and to boost exploitation of the project’s results and its sustainability, I-BiDaaS experts showcased the third and final version of the I-BiDaaS solution and its applicability in 8 real-world, industry-led experiments in the domains of banking, manufacturing, and telecommunication.

In total, 94 attendees from 12 countries attended the event (see Figure 21), a number that significantly increased the visibility of the project and the achieved results.

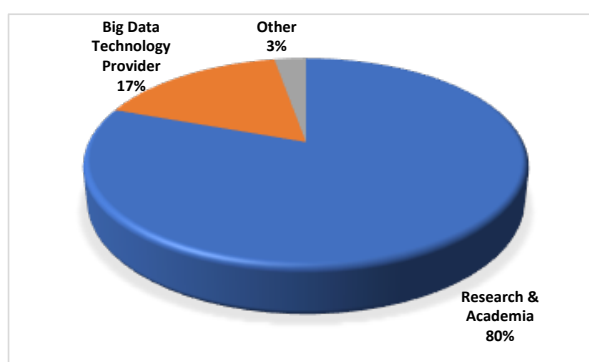
<sup>76</sup> [https://www.big-data-value.eu/wp-content/uploads/2020/06/16.-I-BiDaaS\\_CaixaBank\\_Success\\_story.pdf](https://www.big-data-value.eu/wp-content/uploads/2020/06/16.-I-BiDaaS_CaixaBank_Success_story.pdf)

<sup>77</sup> <https://www.big-data-value.eu/best-success-story-award-2020/>

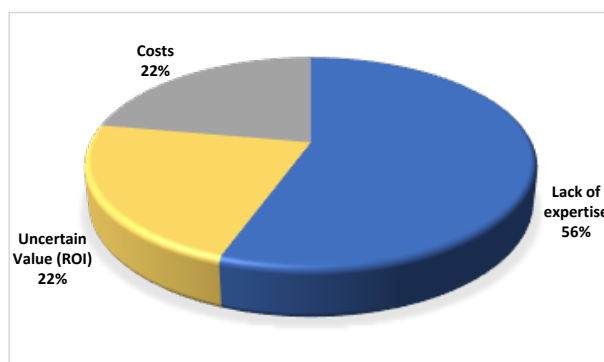


**Figure 21. I-BiDaaS Final Event - Geographical Spread**

Similar to other events organised by the project, the attendees were asked a few questions in order to understand their background and interests. From the responses received, the majority of the attendees were from Research and Academia (81%), followed by Big Data Technology providers (17%), as shown in Figure 22. Even though 53% worked with Big Data and 94% were interested in Big Data technologies to improve their customer experiences. The main barrier to achieve this was considered the lack of expertise (56%), as shown in Figure 23.



**Figure 22. To which of our stakeholder types do you belong?**



**Figure 23. What is the main barrier or risk preventing you from implementing Big Data Analytical solutions in your organisation?**

The event started with the keynote speaker, Nuria de Lama, European Programs Manager, Atos Research & Innovation - Member of Board of Directors, BDVA – I-BiDaaS External Advisory Board member. The title of her talk was ‘A 5-years journey through the European Big Data Landscape’. Nuria de Lama focused on the EU data ecosystem and how it dramatically has evolved in the last years with enormous progress on data analytics and privacy-preserving technologies. She highlighted the challenges that still remain, like making data available and easily usable, agreeing on a European framework for data governance and ensuring the interoperability and openness of data platforms. Questions such as competitiveness and European sovereignty were also part of the discussion. She summarised the main findings, activities and results and opening up a discussion about the future, including the potential suitability of data spaces or infrastructures like GAIA-X to address the aforementioned challenges. The keynote by Nuria de Lama set the ground for the presentations that followed.

After the keynote, an overview of the I-BiDaaS project was presented along with the I-BiDaaS scientific and technical advancements. The three data providers presented the requirements set at the initial phases of the projects and how I-BiDaaS managed to successfully address them. Step by step demos of the I-BiDaaS solution and its application to the targeted sectors was

performed and the key findings and the lessons learned were shared with the audience. The I-BiDaaS platform, with the three different modes, was also showcased together with the two generic use cases that have also been developed. Finally, the I-BiDaaS Innovation Ecosystem and the business and commercial offering of the I-BiDaaS solution was presented.

Interesting questions were received, focusing on data sharing and data availability. More specifically, the attendees were interested to hear the view of the speakers regarding what will be the game changer for convincing data owners to share more data, and given the strict regulations which will be the most prominent way in the next years to introduce new Data Analysis technologies.

**Feedback received from external stakeholders overview.** I-BiDaaS partners organised nine special events (Info Days, Workshops, Hackathons, Webinars, etc.), a number significantly higher than expected. The aim of this strategy was to raise user's participation and awareness, to facilitate stakeholders to adopt and fully leverage the capabilities of the I-BiDaaS platform while also approaching new potential users/customers. Those events provided opportunities for disseminating and exchanging views on the best practices and results based on the I-BiDaaS pilots. From one event to another, I-BiDaaS managed to attract more and more stakeholders who closely followed all the activities of the project. An important pool of stakeholders has been created. Their feedback that was collected during the events or offline where they were invited to participate in the evaluation of the I-BiDaaS solution guided the Consortium's decisions and highly impacted the final version of the I-BiDaaS solution.

### 5.3. Exploitation and potential commercialization

As mentioned in D6.4 [4], spending in the global IT services market will significantly increase by 2021 and onwards. I-BiDaaS outcomes will reach the required maturity peak for exploitation purposes in M36, providing companies with the competitive advantage towards a thriving data-driven EU economy, maximizing the exploitation opportunities for individual partners and the sustainability of the tools in the long-term, beyond the project's lifespan. D7.8 [17] will describe individual exploitation plans for innovations identified by the EU Innovation Radar linking potential I-BiDaaS products to actual market needs, given their capability to address different market stakeholders. I-BiDaaS design allows decoupling a given tool from the rest, making the platform flexible and modular. Modularity has an impact on the pricing model. Therefore, with all consortium partners' participation, the exploitation team designed a methodology for the pricing model, enabling the identification of a suitable strategy and possible synergies between partners and the establishment of an exploitation agreement.

Moreover, the COVID-19 pandemic accelerated the digital transition, which was already reshaping the way companies operate. New technologies such as Big Data Analytics will create new jobs, requiring new and updated skills. A self-service solution like I-BiDaaS intends to empower employees with the right knowledge crucial for long-term and sustainable growth, productivity, innovation, and competitiveness of large, small and medium-sized enterprises (SMEs). Data processing tools and services available in experimental training environments in CaixaBank, CRF, and the internal innovation calls from TID will ensure long-term sustainability and develop a culture of digital skills.

I-BiDaaS will leverage the EU's opportunity to become a global leader in accelerating digital transformation and the purpose of Europe to become a circular industry. To this end, towards the end of the project, the exploitation team did a reality check of the business models and finalized the design and proposition of a sound business plan, including financial modelling, and analysis of the potential revenue streams. The exploitation team also identified possible circular business models in the project use cases, which are considerably insightful for the

different sectors that need a clear plan to boost a range of new skills, including big data analysis. A detailed report of the exploitation activities can be found in D7.8 [17].

## 6. Conclusions

This deliverable provides a concluding assessment of the final I-BiDaaS solution, through a review of a series of real-life industrial experiments from the telecommunication, banking and manufacturing industries, which demonstrate the successful application of the I-BiDaaS solution in real-world environments. All evaluation activities were guided by the I-BiDaaS experimental protocol, described in D6.3 [3], to ensure the successful implementation of the operational experiments and their alignment with the key business objectives, as determined by the industrial partners for the reported use cases.

In more detail, the deliverable reports a comprehensive evaluation of the final I-BiDaaS platform and all of its main components, through a series of exhaustive quantitative and qualitative industry-validated benchmarks that consider both operational (cost, service levels, etc.) and technical terms (performance of solution), as demonstrated by a rigorous use case evaluation, in the context of the three addressed industrial sectors. Furthermore, the assessment of the I-BiDaaS platform is situated beyond the aforementioned use cases, and it is carried out in a wider network of industrial peers and resources, such as the one offered by the European project DataBench, as well as with the support of external stakeholders.

Last, the deliverable provides a critical discussion on the achieved impact (by concluding the work started in D6.2 [2], D6.3 [3] and D6.4 [4]) with respect to the envisioned project innovation, and charts future directions for exploitation and potential commercialization on a global scale. In conclusion, we present the benefits of the participatory evaluation that involved the consortium and external stakeholders in evaluating key results and what constitutes success.

## 7. References

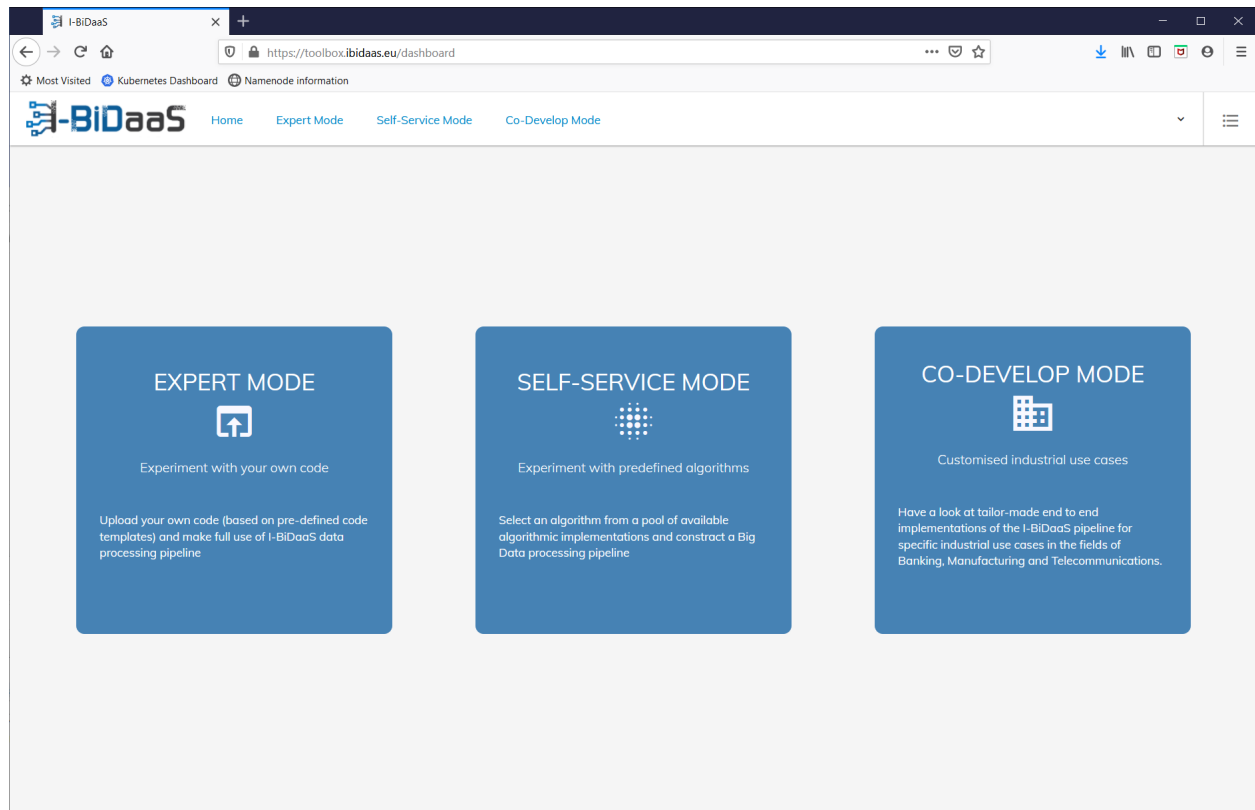
- [1] I-BiDaaS Consortium, Deliverable D1.3: Positioning of I-BiDaaS, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [2] I-BiDaaS Consortium, Deliverable D6.2: Experiments implementation – initial version, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [3] I-BiDaaS Consortium, Deliverable D6.3: Evaluation final report, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [4] I-BiDaaS Consortium, Deliverable D6.4: Experiments implementation (final version), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [5] I-BiDaaS Consortium, Deliverable D5.6: Big-Data-as-a-Self-Service Test and Integration Report (third version), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [6] I-BiDaaS Consortium, Deliverable D6.1: Evaluation report (interim version), Confidential Report
- [7] I-BiDaaS Consortium, Deliverable D3.1: Batch Processing Analytics module implementation, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [8] I-BiDaaS Consortium, Deliverable D3.3: Batch Processing Analytics module implementation (final report), Public Report, available at <https://www.ibidaas.eu/deliverables>
- [9] Álvarez Cid-Fuentes, Javier & Álvarez, Pol & Amela, Ramon & Ishii, Kuninori & Morizawa, Rafael & Badia, Rosa M.. (2019). Efficient development of high performance data analytics in Python. Future Generation Computer Systems. 111. 10.1016/j.future.2019.09.051.
- [10] M. Gupta, J. Gao, C. C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2250-2267, Sept. 2014, doi: 10.1109/TKDE.2013.184.
- [11] I-BiDaaS Consortium, Deliverable D4.1: Real time complex event processing engine – design and approach, Confidential Report, available at <https://www.ibidaas.eu/deliverables>
- [12] I-BiDaaS Consortium, Deliverable D4.2: Distributed event-processing engine, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [13] I-BiDaaS Consortium, Deliverable D4.3: Streaming analytics and predictions, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [14] I-BiDaaS Consortium, Deliverable D7.5: Second report on Dissemination strategy and activities, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [15] I-BiDaaS Consortium, Deliverable D7.3: First report on Dissemination strategy and activities, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [16] I-BiDaaS Consortium, Deliverable D7.7: Third report on Dissemination strategy and activities, Public Report, available at <https://www.ibidaas.eu/deliverables>
- [17] I-BiDaaS Consortium, Deliverable D7.8: Exploitation strategy and activities (third report), Confidential Report, available at <https://www.ibidaas.eu/deliverables>

## ANNEX I – Walk-Through Expert & Self-Service mode

### Walk-Through Expert mode

This document describes the steps required for setting up a new project and running a new experiment in the Expert Mode of the I-BiDaaS toolbox.

First, navigate to <https://toolbox.ibidaas.eu> and login with your credentials. After logging in, select EXPERT MODE (Figure 24).



**Figure 24. Main page**

In the Expert mode screen, select to create a new project (Figure 25).

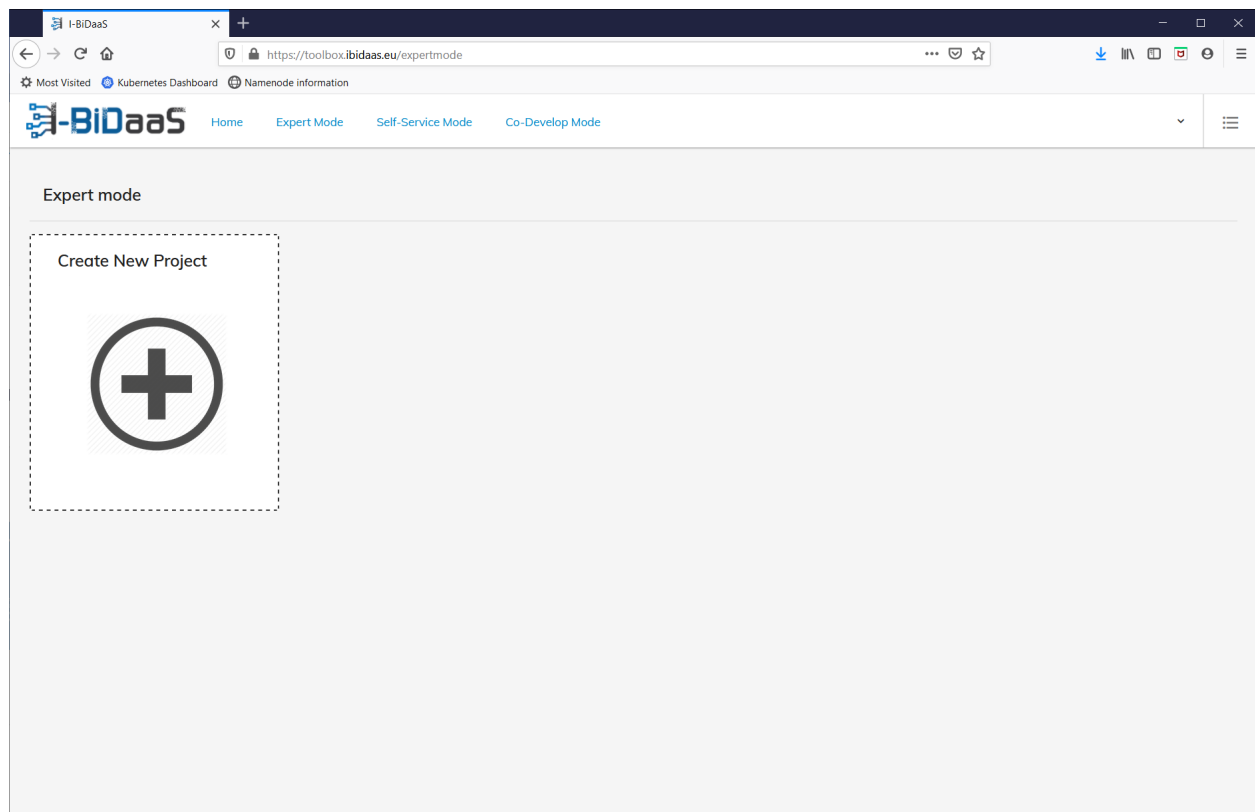


Figure 25. Expert mode

Start adding the new project details. For the new project, enter a name and description, select “None” as input selection and use the default docker image. You can download the `random_forest.zip` code template (Figure 26) and use it exactly as is, or you can choose to modify its contents (the `main.py`) file. In any case, the zip must contain the main python script and optionally any dependencies or even a sample input file.

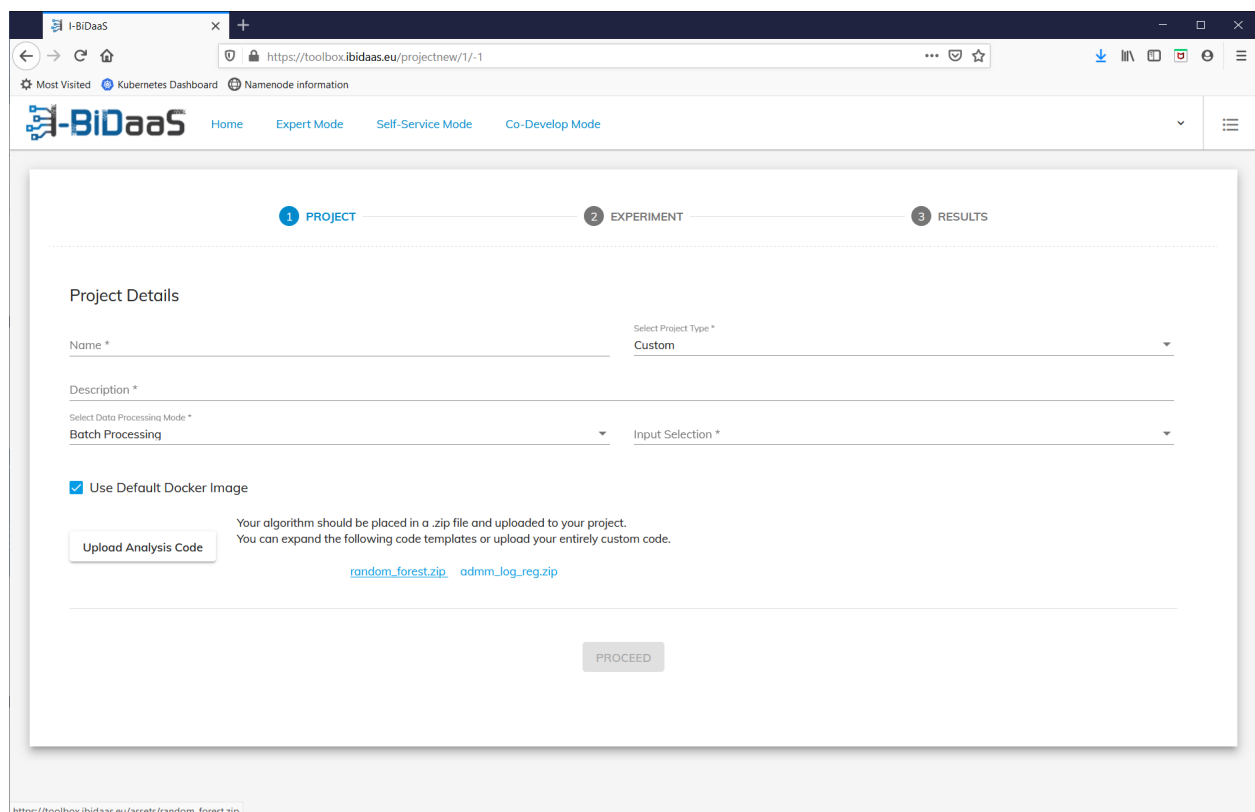
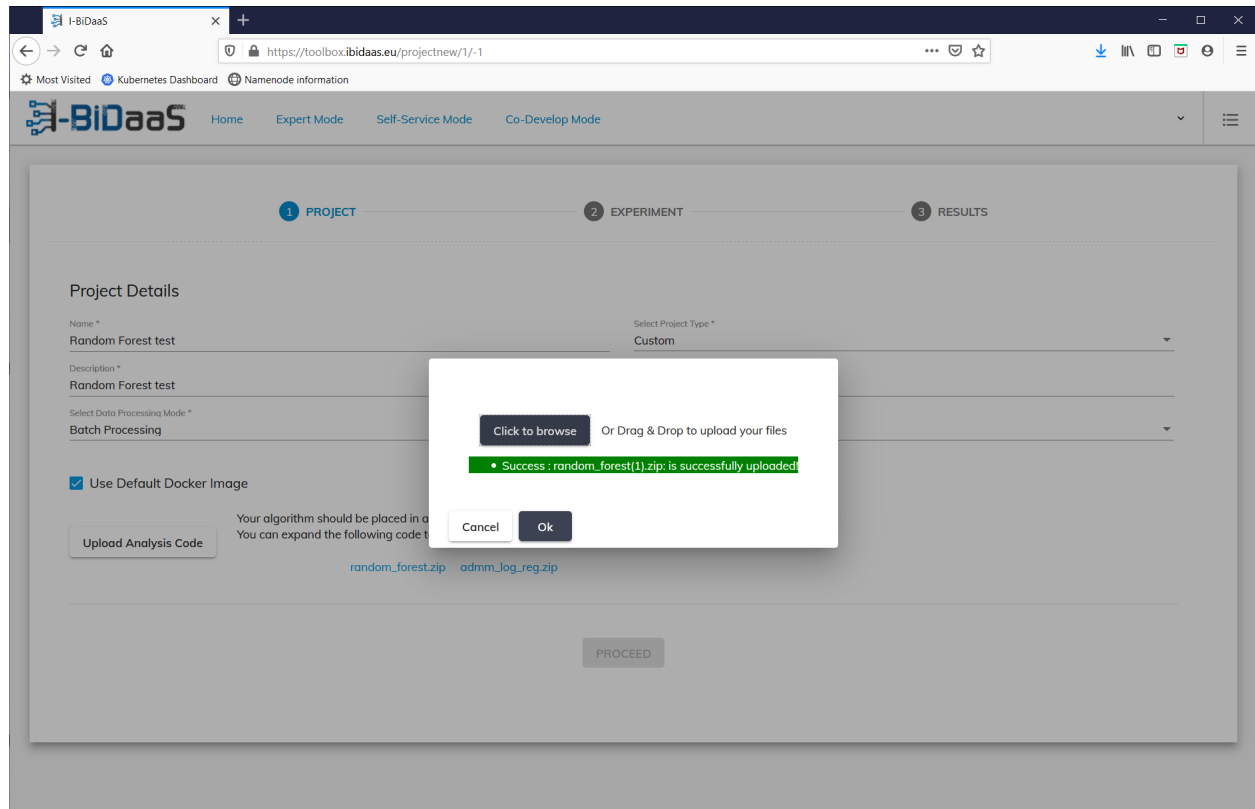


Figure 26. Preparing the project (I)

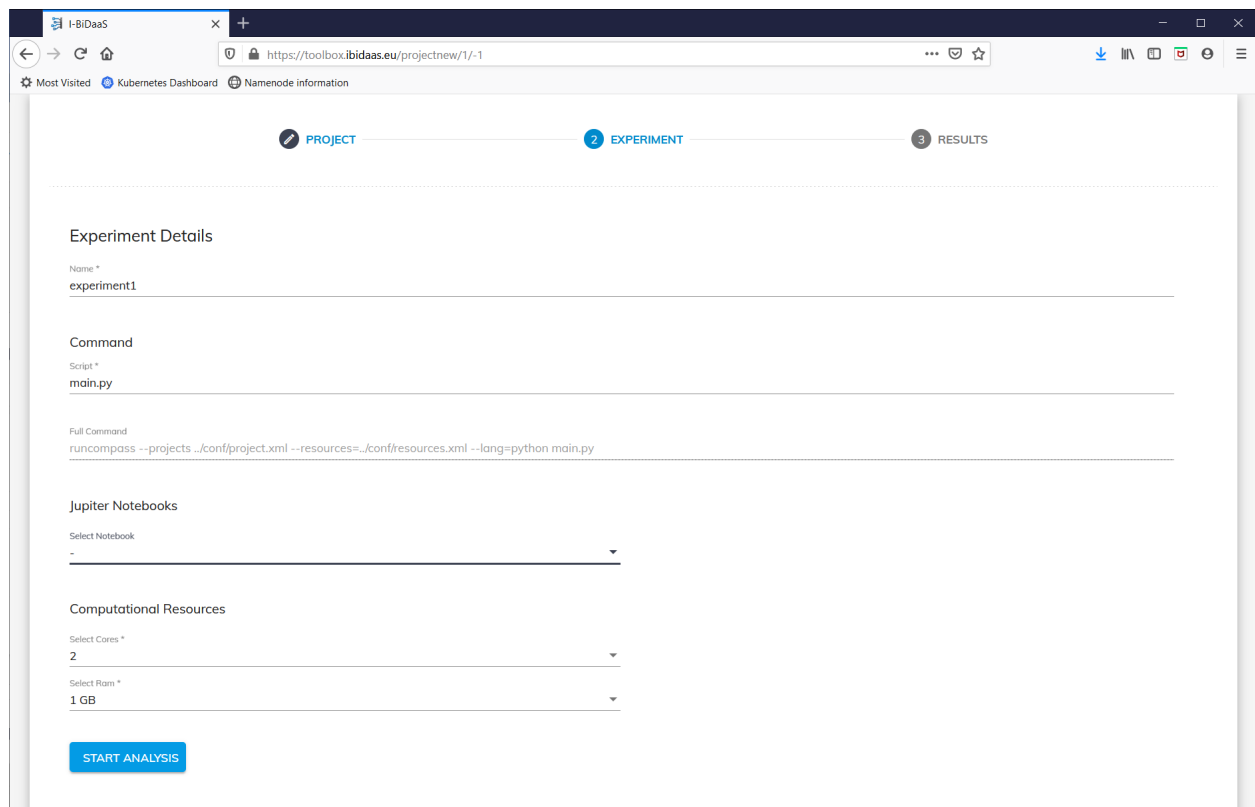


Upload the code as shown in Figure 27. The zip file must not contain any folders.



**Figure 27. Code uploading**

Click “Proceed” and in the next page (Figure 28) set up a new experiment. A project may contain one or more experiments. In each experiment you can modify the command (e.g. provide different parameters) for the same code that was provided during project initialization.



**Figure 28. Experiment details**

After setting up the required parameters of the new experiment you can click “Start Analysis”. The experiment will run for some time (Figure 29) in case of the random forest code, it should be a couple of minutes , depending on the server load).

The screenshot shows the I-BiDaaS web interface in a browser window. The address bar shows the URL <https://toolbox.ibidaas.eu/projectnew/1/-1>. The page has a dark blue header with the I-BiDaaS logo and a progress bar with three steps: PROJECT, EXPERIMENT (active), and RESULTS. The main content area is titled 'Experiment Details' and contains several sections: 'Name' with the value 'experiment1', 'Command' with 'main.py' and a full command line, 'Jupyter Notebooks' with a dropdown menu, and 'Computational Resources' with '2' cores and '1 GB' of RAM. A 'START ANALYSIS' button is at the bottom left.

**Figure 29. Running experiment**

You can leave this page at any time and navigate back the expert mode page and then to the Project details page (Figure 30). There you can see the list of experiments. The new experiment should be in running state for a while.

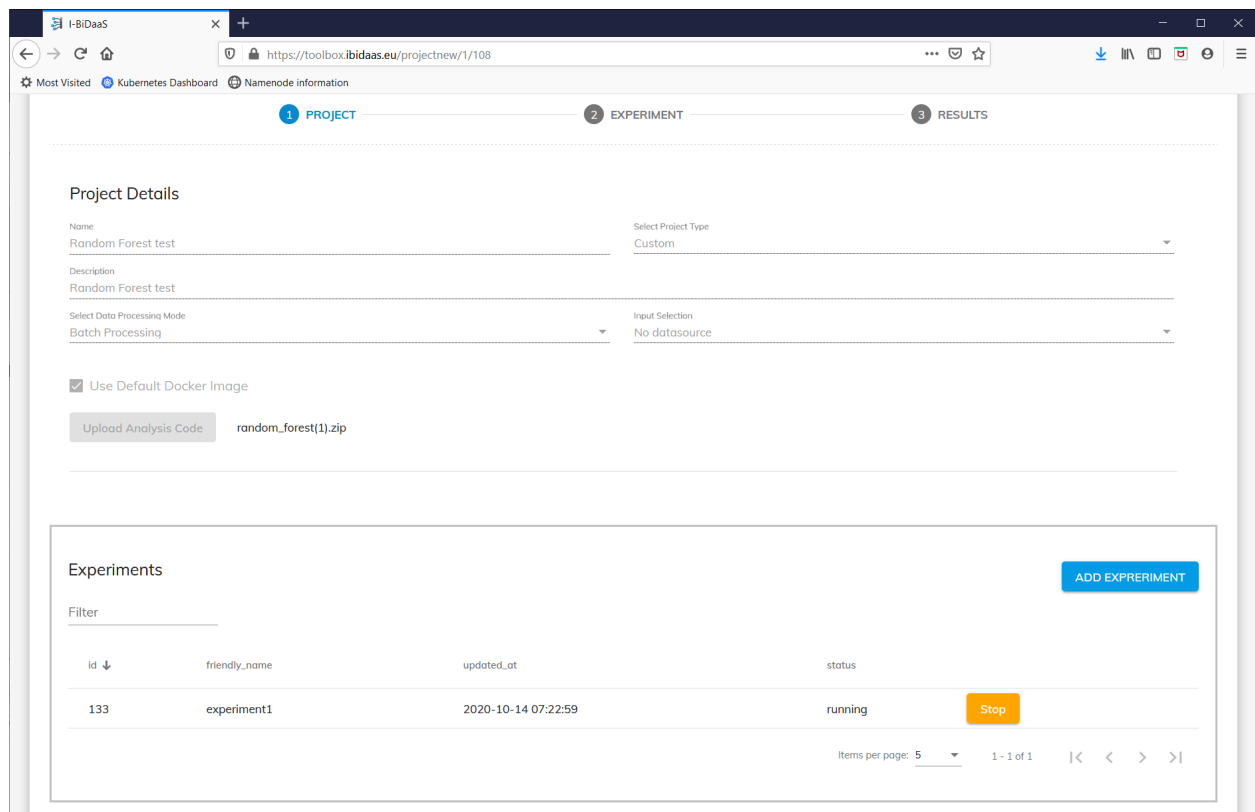


Figure 30. Project details page

When the experiment terminates, the state will be updated to “stopped” and then you can choose to view the results or delete it (Figure 31).

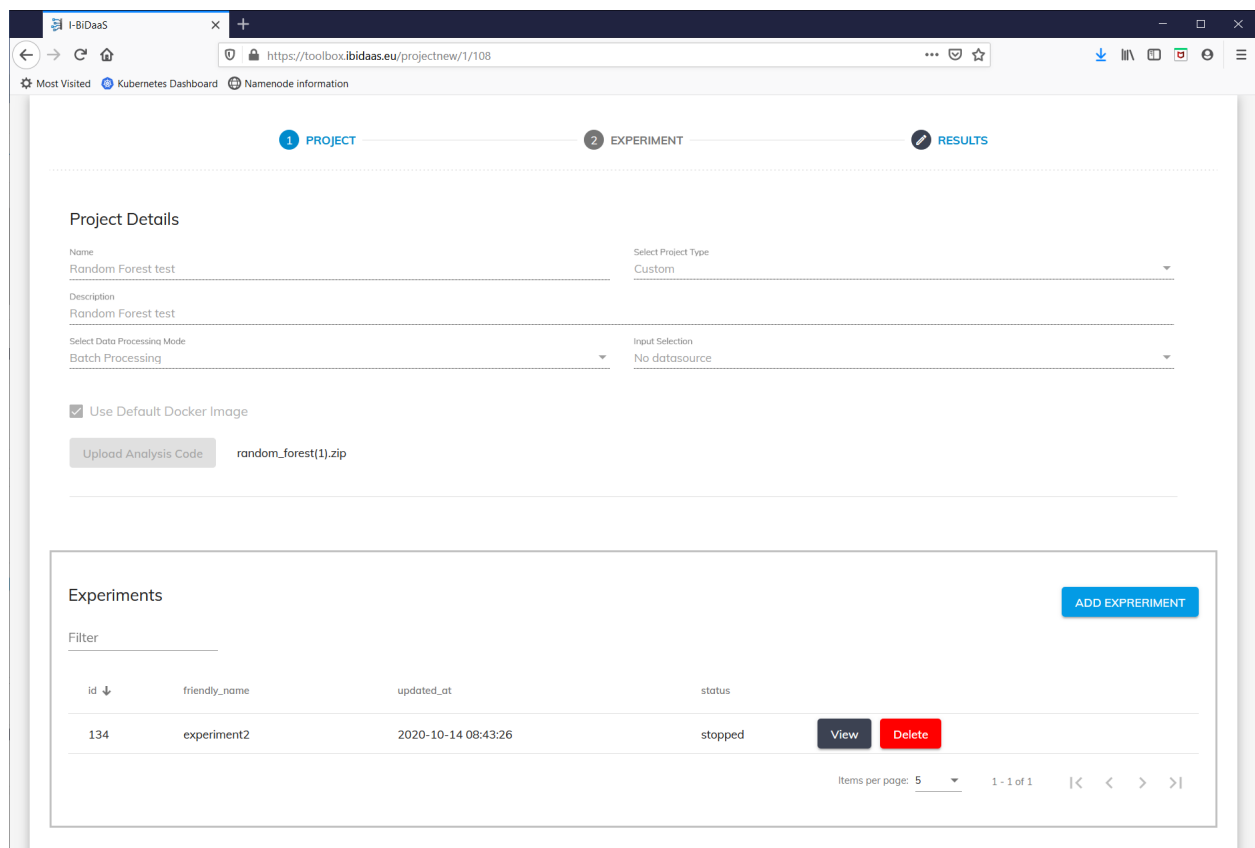
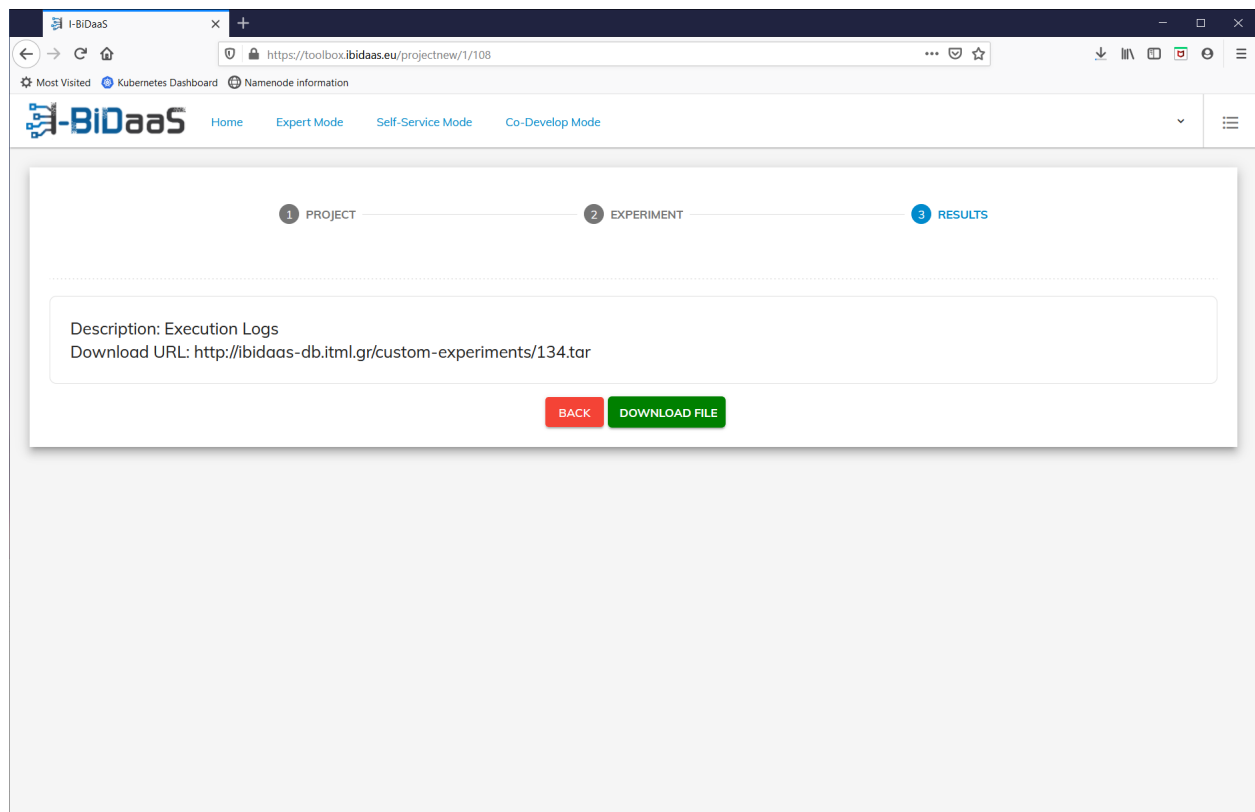


Figure 31. Project details (II)

By selecting “View” you can choose to download a tar ball containing the results (Figure 32).



**Figure 32. Experiment Results page**

The tarball has the following file structure:

- Command.txt: The actual command line used to run the experiment
- Output.txt: Debugging output of the docker swarm. A successful run should look like the screenshot in Figure 33, containing the docker stack creation, execution and removal of the docker stack
- custom-experiment-XXX-results/application\_log.out: the output to stdout of the app. Results could be printed there
- custom-experiment-141-results/application\_log.err: the output of the stderr of the app

Any other file produced by the script will be also included in the file structure.

```

Waiting for resources to become available
192.168.122.15:2376
192.168.122.15:2376
[ RUNCOMPSS-DOCKER ]: Execution summary -----
[ RUNCOMPSS-DOCKER ]: Container cpu units: 2
[ RUNCOMPSS-DOCKER ]: Container memory: 1 GB
[ RUNCOMPSS-DOCKER ]: Image name: ibidaas-db:5000/ibidaas-universal:selfservice
[ RUNCOMPSS-DOCKER ]: Number of workers: 2
[ RUNCOMPSS-DOCKER ]: Swarm manager ip: 192.168.122.15:2376
[ RUNCOMPSS-DOCKER ]: Context directory: /root/general/users/user1/projects/113
[ RUNCOMPSS-DOCKER ]: VM creation time: 60
[ RUNCOMPSS-DOCKER ]: Minimum vms to run: 0
[ RUNCOMPSS-DOCKER ]: Maximum vms to run: 0
[ RUNCOMPSS-DOCKER ]: Stack name: custom-experiment-141
[ RUNCOMPSS-DOCKER ]: -----
[ RUNCOMPSS-DOCKER ]:
[ RUNCOMPSS-DOCKER ]: Generating docker-compose.yml file into '/var/www/custom-experiment-logs/141/custom-experiment-141-compose' ...

[ RUNCOMPSS-DOCKER ]: Executing application in swarm manager...

deploying stack
Creating network custom-experiment-141-docker_runcompss-docker-net
Creating service custom-experiment-141-docker_master
Creating service custom-experiment-141-docker_worker1
Creating service custom-experiment-141-docker_worker2
stack deployed

[ RUNCOMPSS-DOCKER ]: Waiting application to finish...
[ RUNCOMPSS-DOCKER ]: Application finished!
[ RUNCOMPSS-DOCKER ]: Retrieving results from master...
docker host: tcp://192.168.122.15:2376
doing docker cp

[ RUNCOMPSS-DOCKER ]: Results successfully retrieved!

[ RUNCOMPSS-DOCKER ]: Check the application results in './custom-experiment-141-results'
[ RUNCOMPSS-DOCKER ]: In case you had debug enabled, check: './custom-experiment-141-results/debug'

[ RUNCOMPSS-DOCKER ]: Cleaning environment from the execution...

removing
docker stack rm custom-experiment-141-docker
Removing service custom-experiment-141-docker_master
Removing service custom-experiment-141-docker_worker1
Removing service custom-experiment-141-docker_worker2
Removing network custom-experiment-141-docker_runcompss-docker-net
docker stack rm custom-experiment-141-docker
done

[ RUNCOMPSS-DOCKER ]: Waiting some seconds for the deletions to take effect

```

Figure 33. Output.txt contents

In case you chose to run the `random_forest` template code, which loads a predefined dataset provided with the `dislib` library, the classification output will be printed (along with a lot of debugging messages in `application_log.out` (Figure 34).

```

[ (BINDING-COMMONS) ] - @GS_Close_File - COMPSS filename: /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-252
[ (20792) ] API] - Deleting File /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-252
[ (BINDING-COMMONS) ] - @GS_Delete_File - COMPSS filename: /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-252
Predicted values:
  Label Predicted
0      2          1
1      0          0
2      1          1
3      1          1
4      1          1
5      1          1
6      0          0
7      1          1
8      0          0
9      2          2
10     2          2
11     2          2
12     1          2
13     1          1
14     1          1
15     0          0
16     1          1
17     0          0
18     2          2
19     0          0
20     1          1
21     2          2
22     2          2
23     1          1
24     2          2
25     0          0
26     0          0
27     0          0
28     0          0
29     0          0

- Classifier accuracy: 0.9333333333333333
[ (20821) ] API] - Deleting File /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-305
[ (BINDING-COMMONS) ] - @GS_Delete_File - COMPSS filename: /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-305
[ (20825) ] API] - Deleting File /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-91
[ (BINDING-COMMONS) ] - @GS_Delete_File - COMPSS filename: /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-91
[ (20827) ] API] - Deleting File /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-280
[ (BINDING-COMMONS) ] - @GS_Delete_File - COMPSS filename: /root/.COMPSS/main.py_01/tmpFiles/pycompssr9rlyk1l/compss-serialized-obj_66c16332-0e0b-11eb-b807-02420a000605-280

```

Figure 34. Sample of the successful execution of random forest template

**Existing code templates:**

Name	Description	Executable script	Command to run
<b>Random forest</b>	A random forest implementation based on dislib <sup>78</sup> , the example code loads a predefined dataset stored within the library. It prints the predicted values in the stdout (application_log.out).	main.py	main.py
<b>ADMM Logistic Regression</b>	A logistic regression algorithm. The example code runs with 2 workers and the dataset provided with the code. It prints the dimensions of the dataset and the total elapsed time.	Main_LogReg_ADMM.py	Main_LogReg_ADMM.py 2

---

<sup>78</sup> <https://www.bsc.es/research-and-development/software-and-apps/software-list/dislib>

## Walk-Through Self-Service mode

This document describes the steps required for setting up a new project and running a new experiment in the Self-Service Mode of the I-BiDaaS toolbox.

First, navigate to <https://toolbox.ibidaas.eu> and login with your credentials. After logging in, select SELF-SERVICE MODE (Figure 35).

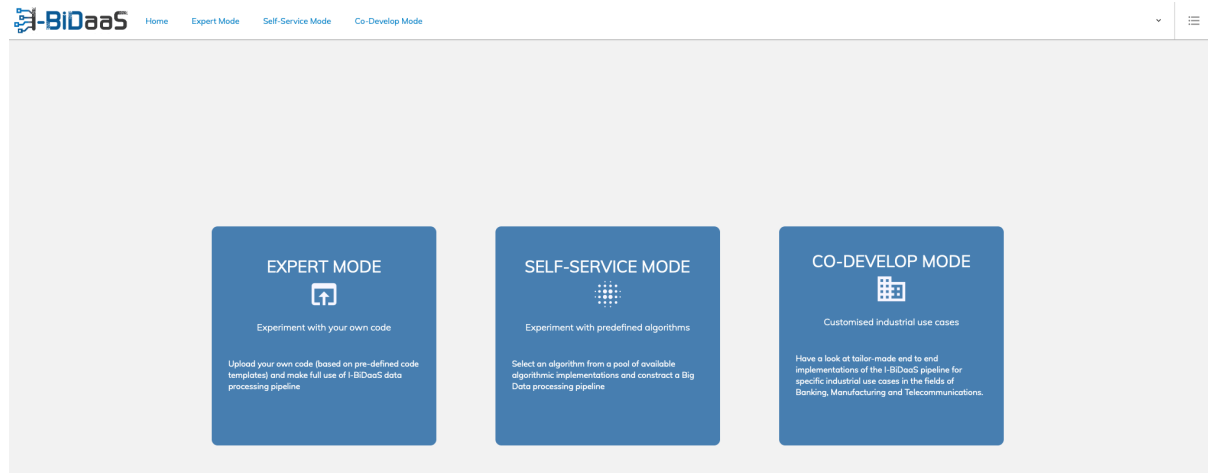


Figure 35. I-BiDaaS main page

In the Self-Service mode screen, you will find several available algorithms (Figure 36) but during this evaluation, we will focus on the following:

- LASSO ADMM
- K-Means – Prediction
- K-Means - Evaluation

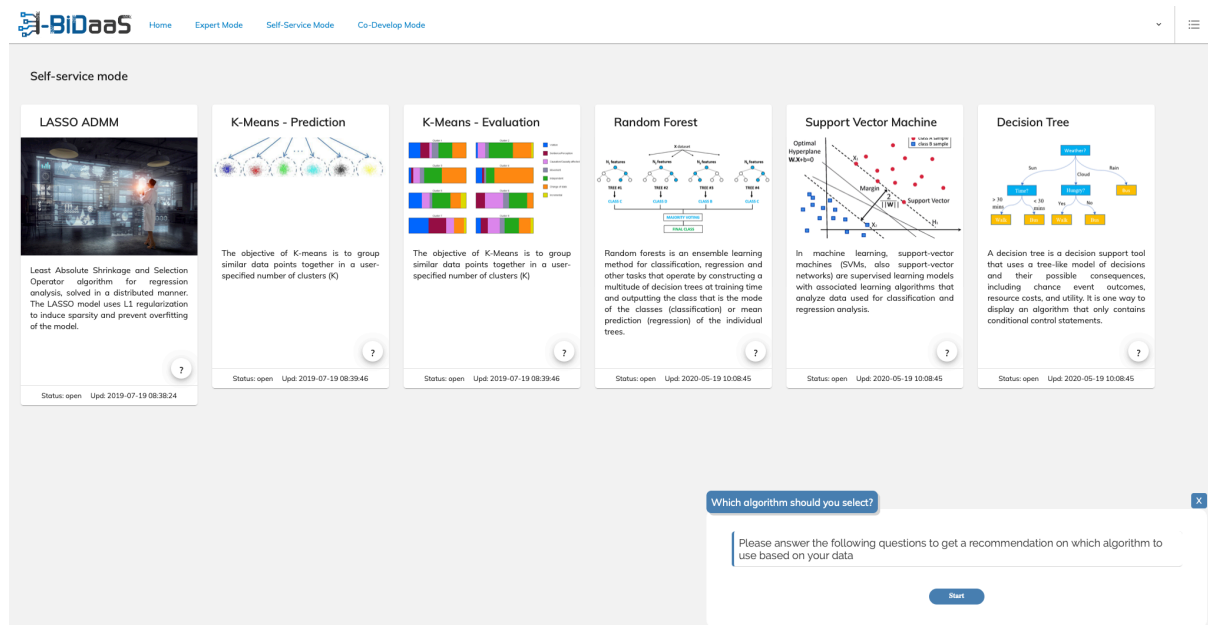
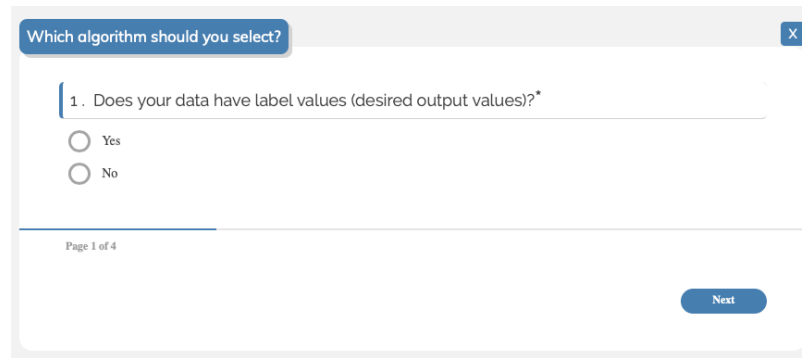


Figure 36. Self-Service Mode page

Before proceeding with the specific experiments, take a look at the questionnaire. Its main goal is to help you get a recommendation on which algorithm you should use based on your data! Do you find this useful? (Figure 37)



When you have finished, click “x” and the questionnaire will be minimized at the right bottom of your screen.

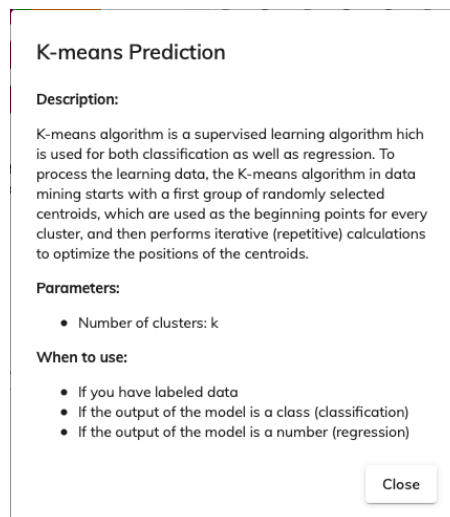


**Figure 37. I-BiDaaS questionnaire for selecting the appropriate algorithm**

Ready to start?

## Experiment #1 – K-means Prediction

Do you want to have a quick look at this algorithm before executing? Click the “?” and a pop-up window will appear on your screen (Figure 38).



**Figure 38. K-means Prediction – at a glance**

K-means Prediction algorithm can be used for clustering a dataset with scalar attributes. Click on the ‘*K-means Prediction algorithm*’ box and you will be redirected to the Project Details page (Figure 39).

The screenshot shows the 'PROJECT' tab selected in the top navigation bar. The page is titled 'Project Details' and contains the following fields:

- Name:** K-Means - Prediction
- Select Project Type:** Preconfigured
- Description:** The objective of K-means is to group similar data points together in a user-specified number of clusters (K)
- Select Data Processing Mode:** Batch Processing
- Input Selection:** File Input
- ☒ Use Default Docker Image

Below the project details is an 'Experiments' section with a table listing existing experiments:

id	friendly_name	updated_at	status	
516	K test 3	2019-08-26 08:59:25	stopped	<a href="#">View</a> <a href="#">Delete</a>
471	New Experiment Kmeans Test	2019-08-16 15:00:31	stopped	<a href="#">View</a> <a href="#">Delete</a>
414	Test experiment on project 1	2019-08-09 12:55:22	stopped	<a href="#">View</a> <a href="#">Delete</a>

At the bottom right of the experiments table, there is a pagination control showing 'Items per page: 5' and '1 - 3 of 3'.

**Figure 39. K-means Prediction – Project Details page**

In order to run a new experiment, click 'Add Experiment' button.

The screenshot shows the 'EXPERIMENT' tab selected in the top navigation bar. The page is titled 'Experiment Details' and contains the following fields:

- Name \***
- Datasource Name \***
- Parameters:**
  - Number of clusters (K):** 10
  - Maximum number of iterations prior to automatic termination:** 10
  - Tolerance (used to compute the threshold for early termination of the algorithm):** 0.0001
  - Chunk size of each chunk (concerns the splitting of our data into chunks, not K-Means directly):** 300
  - Portion of data to be treated as the test data (remaining data will be used for training):** 0.2
- Computational Resources:**
  - Select Cores \*:** 2
  - Select Ram \*:** 1 GB

**Figure 40. K-means Prediction – Experiment details**

Here you can set up the required parameters of the new experiment (Figure 40). The available dataset to be selected is `/root/general/fabricated/kmeans-prediction-data.csv`, which is a csv file with coordinates. As soon as you select the dataset, you will be able to preview a portion of the dataset, as illustrated in Figure 41.

Experiment Details

Name \*

Test\_step-by-step

Datasource Name \*

/root/general/fabricated/kmeans-prediction-data.csv

CSV Preview

5.927870038042057	-3.7605720644165266
3.6671797411710543	-3.5827759470158864
6.137837503975408	-2.5597718857602305
6.407218966021258	-2.4207235749246476
7.798664471833868	3.7016246674039843
7.505601361167013	3.7875310467710954

**Figure 41. K-means Prediction – data preview**

Modify the parameters of the algorithm (number of clusters, max number of iterations, etc.). Before you start the analysis do not forget to select the desired Computational Resources (Cores, Ram) as depicted in Figure 42.

Computational Resources

Select Cores \*

4

Select Ram \*

8 GB

**START ANALYSIS**

**Figure 42. K-means Prediction – select Computational Resources**

Ready? Click ‘*Start Analysis*’. The experiment will run for some time (Figure 43).

6.137837503975408 -2.5597718857602305

6.407218966021258 -2.4207235749246476

7.798664471833868 3.7016246674039843

7.505601361167013 3.7875310467710954

Parameters

Number of clusters (K)

10

Maximum number of iterations prior to automatic termination

10

Distance used to compute the threshold for early termination of the algorithm

0.0001

Chunk size of each chunk (controls the splitting of our data into chunks, not K-Means directly)

300

Portion of data to be treated as the test data (remaining data will be used for training)

0.2

Computational Resources

Select Cores \*

4

Select Ram \*

8 GB

**START ANALYSIS**

**Figure 43. K-means Prediction – Running experiment**

You can leave this page at any time and navigate back to the self-service mode page and then to the Project details page (Figure 44). There you can see the list of experiments that you have executed. The new experiment should be in running status for a while.

**Project Details**

Name: K-Means - Prediction

Select Project Type: Preconfigured

Description: The objective of K-means is to group similar data points together in a user-specified number of clusters (K)

Select Data Processing Mode: Batch Processing

Input Selection: File Input

☒ Use Default Docker Image

---

**Experiments** ADD EXPERIMENT

Filter

id ↓	friendly_name	updated_at	status	
1906	Test_step-by-step	2020-11-10 14:57:51	running	<span>Stop</span>
516	K test 3	2019-08-26 08:59:25	stopped	<span>View</span> <span>Delete</span>

**Figure 44. K-means Prediction – Project details page**

When the experiment terminates, the status will be updated to *'stopped'* and then you can choose to view the results or delete it (Figure 45).

**Project Details**

Name: K-Means - Prediction

Select Project Type: Preconfigured

Description: The objective of K-means is to group similar data points together in a user-specified number of clusters (K)

Select Data Processing Mode: Batch Processing

Input Selection: File Input

☒ Use Default Docker Image

---

**Experiments** ADD EXPERIMENT

Filter

id ↓	friendly_name	updated_at	duration	status	
1940	Test_step-by-step	2020-11-24 11:23:44	00:37:19	stopped	<span>View</span> <span>Delete</span>

**Figure 45. K-means Prediction – Project details**

Click *'View'* and then *'Proceed'* to see the results. You can also download the results (Figure 46).



Figure 46. K-means Prediction – Visualise results

Repeat the same experiment several times by changing the algorithm's input parameters and compare the results!

## Experiment #2 – K-means Evaluation

Another available algorithm in the I-BiDaaS Self-Service mode is K-Means Evaluation that can be used for training a K-means classifier, using a labelled dataset of scalar attributes.

Click the '*K-means Evaluation algorithm*' box and you will be redirected to the Project Details page (Figure 47).

**Project Details**

Name: K-Means - Evaluation

Select Project Type: Preconfigured

Description: The objective of K-Means is to group similar data points together in a user-specified number of clusters (K)

Select Data Processing Mode: Batch Processing

Input Selection: File Input

☒ Use Default Docker Image

**Experiments**

Filter

id	friendly_name	updated_at	status		
1726	Test_JN	2020-05-29 12:54:32	stopped	View	Delete
1722	Test_sf	2020-05-29 10:23:11	stopped	View	Delete
1705	clusters2	2020-05-15 13:43:12	stopped	View	Delete
1704	ewrver	2020-05-15 13:36:09	stopped	View	Delete

[ADD EXPERIMENT](#)

**Figure 47. K-means Evaluation – Project Details page**

In order to run a new experiment, click ‘*Add Experiment*’ button.

#### Experiment Details

Name \*

Test\_step-by-step

Datasource Name \*

#### Parameters

Number of clusters (K)

2

Maximum number of iterations prior to automatic termination

5

Tolerance (used to compute the threshold for early termination of the algorithm)

0.0001

Portion of data to be treated as the test data (remaining data will be used for training)

0.2

Chunk size of each chunk (concerns the splitting of our data into chunks, not K-Means directly)

300

#### Computational Resources

Select Cores \*

4

Select Ram \*

4 GB

**Figure 48. K-means Evaluation – Experiment details**

Set up the required parameters of the new experiment (Figure 48) as before. The available dataset to be selected is `/root/general/fabricated/kmeans-evaluation-data.csv`, which is a labelled csv file with coordinates and the cluster id. Number of labels: 10 , values [0 – 9]

As soon as you select the dataset, you will be able to preview a portion of the dataset, as illustrated in Figure 49.

## Experiment Details

Name *		
Test_step-by-step		
Datasource Name *		
/root/general/fabricated/kmeans-evaluation-data.csv		
CSV Preview		
5.927870038042057	-3.7605720644165266	5
3.6671797411710543	-3.5827759470158864	3
6.137837503975408	-2.5597718857602305	6
6.407218966021258	-2.4207235749246476	6
7.798664471833868	3.7016246674039843	7
7.505601361167013	3.7875310467710954	7

Figure 49. K-means Evaluation – data preview

Modify the parameters of the algorithm (number of clusters, max number of iterations, etc.). Before you start the analysis do not forget to select the desired Computational Resources (Cores, Ram).

Ready? Click ‘*Start Analysis*’ The experiment will run for some time. You can leave this page at any time and navigate back to the self-service mode page and then to the Project details page (Figure 50). There you can see the list of experiments that you have executed. The new experiment should be in running status for a while.

**Project Details**

Name: K-Means - Evaluation

Select Project Type: Preconfigured

Description: The objective of K-Means is to group similar data points together in a user-specified number of clusters (K)

Select Data Processing Mode: Batch Processing

Input Selection: File Input

☒ Use Default Docker Image

**Experiments**

id	friendly_name	updated_at	status	
1909	Test_step-by-step	2020-11-10 17:39:41	running	Stop
1726	Test_JN	2020-05-29 12:54:32	stopped	View Delete

ADD EXPERIMENT

Figure 50. K-means Evaluation – Project details page

When the experiment terminates, the status will be updated to ‘*stopped*’ and then you can choose to view the results or delete it (Figure 51).



1 PROJECT

2 EXPERIMENT

3 RESULTS

---

Project Details

Name

K-Means - Evaluation

Select Project Type

Preconfigured

Description

The objective of K-Means is to group similar data points together in a user-specified number of clusters (K)

Select Data Processing Mode

Batch Processing

Input Selection

File Input

☒ Use Default Docker Image

---

Experiments

Filter

ADD EXPERIMENT

id	friendly_name	updated_at	duration	status	
1942	Test_step_by_step	2020-11-24 11:41:23	00:00:34	stopped	<div>View</div> <div>Delete</div>

Figure 51. K-means Evaluation – Project details page

Click ‘View’ and then ‘Proceed’ to see the results. You can also download the results (Figure 52).



### Figure 52. K-means Evaluation – Visualise results

Repeat the same experiment several times by changing the algorithm's input parameters and compare the results!

## Experiment #3 – ADMM Lasso

Now that you are feel more comfortable with the Self-Service mode of the I-BiDaaS platform, try our third and final experiment with ADMM Lasso algorithm (Figure 53)!

Lasso

Description:

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Parameters:

- Constraint on the sum of the absolute values of the model parameters:  $\lambda$

When to use:

- If you have labeled data
- If the output of the model is a number (regression)
- When your dataset is high-dimensional

Close

Figure 53. ADMM Lasso – Quick preview

ADMM LASSO (Least Absolute Shrinkage and Selection Operator) Lasso – a sparse regression model: Explain your output variable through a linear model of your features, using a small number of features.

Select '*Add Experiment*' and fine tune the parameters as you did with the previous experiments. This time the data source to be selected is: `/root/general/fabricated/admm_lasso`.

In case you are interested for more information regarding ADMM:

### What is it?

- ADMM – Alternating Direction Method of Multipliers: a generic optimization solver for parallel/distributed systems

### Why do I need it?

- Train various supervised and unsupervised ML models
- (e.g., Logistic regression, Support vector machines, Clustering, etc.)

### Why should I use it?

- Robustness (e.g., provably works under arbitrary step-size, as opposed to, e.g., SGD or BFGS)
- Widely applicable (e.g., for non-smooth problems like SVM where others may fail)
- Minor adaptation needed for a completely different ML model
- Parallelizes work over arbitrary number of nodes/cores

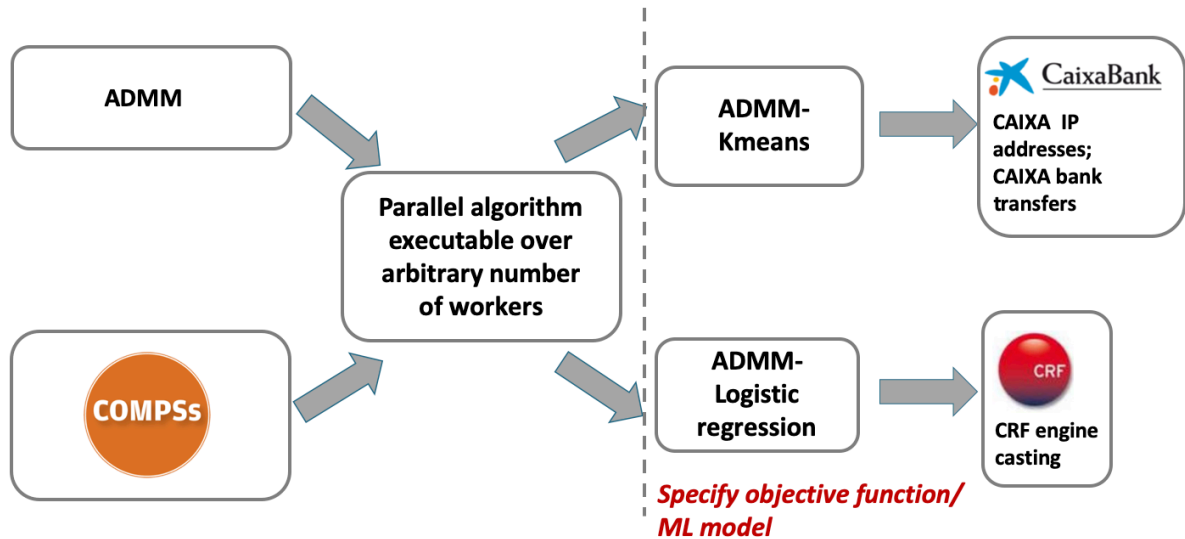


Figure 54. ADMM in I-BiDaaS – How it works and a usage example –

### Tune parameters through the user interface (Figure 55)

- Penalty parameter – how sparse your solution will be; e.g., do you want to explain your output with a very few features
- # Workers over which to parallelize – how fast you want your solution
- Absolute and relative tolerance – how accurate you want your solution
- **Set the amount of computational resources**
- # Cores
- RAM memory size
- **Visualize results (Figure 56)**
- Primal and dual residual – how well your model explains the data/how close you are to the “ideal” solution

Figure 55. ADMM Lasso – Project details page

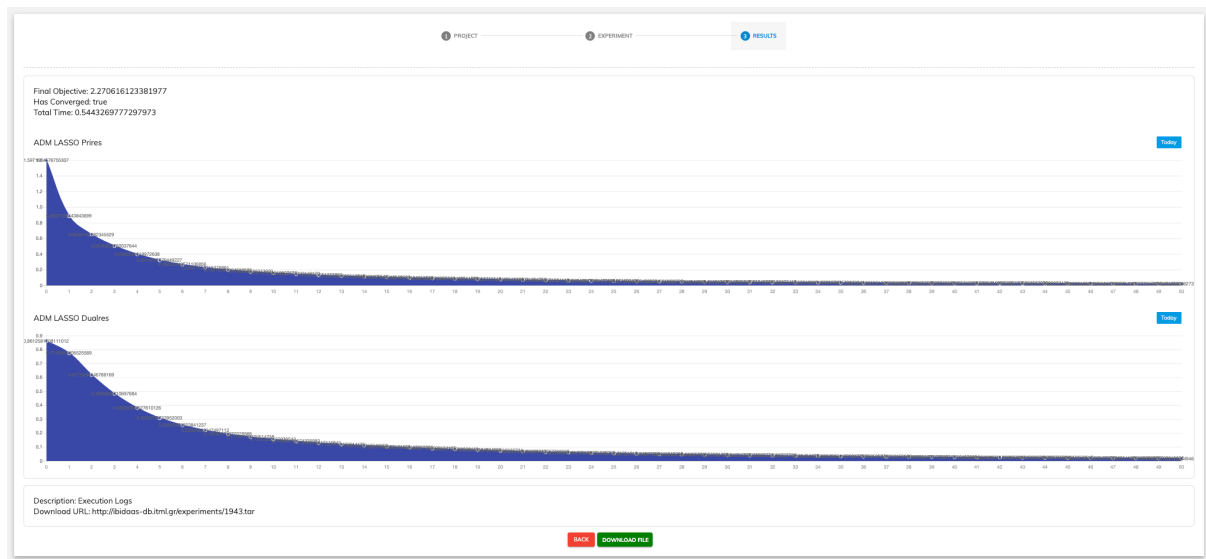


Figure 56. ADMM Lasso – Visualise results

***The evaluation of the Self-Service Mode of the I-BiDaaS Solution is completed! Your feedback is highly appreciated! Do not forget to fill our questionnaire!***

## ANNEX II – Standardisation activities

**Table 34: List of Standards used in I-BiDaaS Technologies**

Service/component	Partner	Standards used in I-BiDaaS
<b>GPU accelerator technology</b>	FORTH	JSON
<b>COMPSs (Programming Model and Runtime)</b>	BSC	ssh, XML, JSON
<b>Hecuba</b>	BSC	TCP Sockets
<b>QBeast</b>	BSC	JSON, WebSocket, ZMQ (de-facto)
<b>Test Data Fabrication</b>	IBM	HTTP, URI, XML, JSON, SQL
<b>Apama Streaming Analytics Platform</b>	SAG	MQTT, JDBC, JSON
<b>Universal Messaging</b>	SAG	JMS, MQTT 3, AMQP 1
<b>Pool of ML algorithms based on structured (non)convex optimization</b>	UNSPMF	ssh, XML, JSON, inherited from COMPSs
<b>Orchestrator</b>	ITML	HTTP/HTTPS JSON
<b>Advanced visualization and monitoring</b>	AEGIS	HTTP/HTTPS, JSON, WebSocket
<b>Resource management and orchestration module – Adaptation Engine Submodule</b>	ATOS	JSON, TOSCA
<b>Resource management and orchestration module – Cloudify Cloud Orchestrator Submodule</b>	ATOS	JSON, TOSCA
<b>MashZone</b>	SAG	JDBC

**Table 35: List of Standards used in the use-case experiments**

No.	Use Case	Related Standards
1	Accurate location prediction with high traffic and visibility	<ul style="list-style-type: none"> <li>General programming, information exchange and information system design: Python, C/C++, XML, SQL, Java EE, HTML5</li> </ul>
2	Optimization of placement of telecommunication equipment	<ul style="list-style-type: none"> <li>General programming, information exchange and information system design: Python, C/C++, XML, SQL, Java EE, HTML5</li> </ul>
3	Quality of Service in Call Centers	<ul style="list-style-type: none"> <li>General programming, information exchange and information system design: Python, C/C++, XML, SQL, Java EE, HTML5</li> <li>Speech data collection, annotation and meta-information: as defined by international organizations such as LDC, NIST, ELRA, like as current standards for speech transcription formats: STM and CTM.</li> <li>General quality assurance system for manufacturing and service industries (TID): ISO 9001, IQNET</li> <li>Environmental protection (TID): ISO 14001</li> </ul>
4	Enhance control of customers to online banking access	<ul style="list-style-type: none"> <li>General programming, information exchange and information system design: Python, JSON, SQL, WebSockets, HTML5</li> <li>BCBS 239 is the Basel Committee on Banking Supervision's standard number 239 (Principles for effective risk data aggregation and risk reporting).</li> <li>ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements.</li> </ul>

		<ul style="list-style-type: none"> <li>• ISO/IEC 27031 — Guidelines for information and communication technology readiness for business continuity.</li> <li>• ISO/IEC 27001 — Information technology - Security Techniques - Information security management systems — Requirements</li> </ul>
5	Advanced analysis of bank transfer payment in financial terminal	<ul style="list-style-type: none"> <li>• General programming, information exchange and information system design: Python, JSON, SQL, WebSockets, HTML5</li> <li>• BCBS 239 is the Basel Committee on Banking Supervision's standard number 239 (Principles for effective risk data aggregation and risk reporting).</li> <li>• ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements.</li> <li>• ISO/IEC 27031 — Guidelines for information and communication technology readiness for business continuity.</li> <li>• ISO/IEC 27001 — Information technology - Security Techniques - Information security management systems — Requirements</li> </ul>
6	Analysis of relationships through IP addresses	<ul style="list-style-type: none"> <li>• General programming, information exchange and information system design: Python, JSON, SQL, WebSockets, HTML5</li> <li>• BCBS 239 is the Basel Committee on Banking Supervision's standard number 239 (Principles for effective risk data aggregation and risk reporting).</li> <li>• ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements.</li> <li>• ISO/IEC 27031 — Guidelines for information and communication technology readiness for business continuity.</li> <li>• ISO/IEC 27001 — Information technology - Security Techniques - Information security management systems — Requirements</li> </ul>
7	Maintenance and monitoring of production assets	CRF Internal procedures
8	Production process of aluminium die-casting	CRF Internal procedures

Table 36: List of Standards per layer related to I-BiDaaS

Service layer functional components		
Functional Components	Component	Standards related to I-BiDaaS
Data Collection	Workstation (Internal system) - CRF	CRF internal procedures
	TID	Speech data collection as defined by international organizations such as LDC, NIST, ELRA, like as current standards for speech transcription formats: STM and CTM.
	CAIXA	CAIXA internal procedures ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements. ISO/IEC 27031 — Guidelines for information and communication technology readiness for business continuity.
	AVT	HTTP/HTTPS, JSON, WebSocket

<b>Data Visualisation</b>	TID	General programming, information exchange and information system design: Python, C/C++, XML, SQL, Java EE
<b>Data pre-processing</b>	TID	Speech data annotation as defined by international organizations such as LDC, NIST, ELRA, like as current standards for speech transcription formats: STM and CTM.
	CAIXA	General programming, information exchange and information system design: Python, JSON.
<b>Data analysis</b>	Manual Statistical analysis - CRF	Automated analysis
	CAIXA	General programming, information exchange and information system design: Python, JSON.
	TID	General programming, information exchange and information system design: Python, C/C++, XML, SQL, Java EE, HTML5
<b>Data Storage</b>	<b>CRF server (storage system)</b>	File storage (CSV, JPG)
	<b>CAIXA</b>	File storage (CSV), Database storage (Oracle DB)
<b>Resource layer functional components</b>		
<b>Distributed processing</b>	COMPSs + Hecuba + Qbeast - BSC	See above. Deployment and remote execution through containerisation (Docker, de-facto standard)
<b>Resource layer functional components</b>		
<b>Integration</b>		
<b>Security systems</b>	FCA security system - CRF	FCA standard
	TID	Environmental protection (TID): ISO 14001
	CAIXA	ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements ISO/IEC 27001 — Information technology - Security Techniques - Information security management systems — Requirements

**Table 37: Analysis of standard processes and methodologies per use-case**

Partner	Methodologies and processes
CAIXA	<ul style="list-style-type: none"> <li>• Data storage and ingestion from various data sources and its preparation</li> <li>• Fabrication of synthetic data with realistic properties for experimentation and testing</li> <li>• Real data anonymisation</li> <li>• Data cleaning</li> <li>• Batch and stream analytics for increasing the speed of data analysis</li> <li>• Visualisation of results and interaction capabilities for the end-users</li> </ul>
TID	<ul style="list-style-type: none"> <li>• Data storage and ingestion from various data flows</li> <li>• Fabrication of synthetic data with realistic properties for experimentation and testing</li> <li>• Data cleaning</li> <li>• Data anonymization</li> </ul>



	<ul style="list-style-type: none"> <li>Batch and stream analytics (including predictive modeling) for increasing the speed of data analysis</li> <li>Visualisation of results and interactive dashboards for the end-users</li> </ul>
CRF	<ul style="list-style-type: none"> <li>Data storage and ingestion from various data sources and its preparation</li> <li>Fabrication of realistic synthetic data for experimentation and testing</li> <li>Real data anonymisation</li> <li>Data cleaning</li> <li>Batch and stream analytics for increasing the speed of data analysis</li> <li>Visualisation of results and interaction capabilities for the end-users</li> </ul>

**Table 38: Participation to Standardisation Bodies per partner**

Partner	Participation to Standardisation bodies
<b>FORTH</b>	BDVA associate member ENISA, first.org
<b>BSC</b>	BDVA member (full member) RDA Alliance (RDA Spain Coordinators, administrative role)
<b>IBM</b>	BDVA full member BDVA, OASIS, W3C, ANSI, SNIA, IEEE, ITU...
<b>SAG</b>	BDVA full member DMG member which defines PMML
<b>ATOS</b>	BDVA full member, FIWARE, GAIA-X, AI4EU, ALASTRIA, EBSI, NESSI, ETSI, ARTEMIS, EARPA, IDS, EFFRA, ETSI, MANO NFV
<b>ITML</b>	BDVA associate member
<b>UNSPMF</b>	BDVA associate member
<b>CRF</b>	FCA Standard
<b>CAIXA</b>	First.org, FI-ISAC, ALASTRIA, EBSI, ESBG Cloud Certification Working Group, ENISA's Financial experts Working Group, EPC Payment Security Support Group and Card Fraud Prevention (EPC) International WG.
<b>TID</b>	3GPP Rel-16 O-RAN Alliance

## Cloud computing – Functional architecture of big data as a service

**Table 39: Mapping between requirements, activities and functional components**

Requirements in [ITU-T Y.3600]	Activities in [ITU-T Y.3600]	Functional components in this Recommendation	Related layers with [ITU-T Y.3502]	Notes
<Clause 8.1 requirement (4)> It is recommended for CSN: data provider (DP) to provide a brokerage service to CSP:BDIP for searching accessible data.	Brokerage data (7.1.1.3)	Data collection functional component (7.1.1)	Service layer	Big Data Infrastructure made available within I-BiDaaS through the Cloud Service Partner in which data collected from sensors are stored according to the data type

<Clause 8.1 requirement (1)> It is required for the CSP:BDIP to support collecting data from multiple CSN: DPs in parallel.	Perform data collection (7.1.3.1)	Data collection functional component (7.1.1) Hecuba UM as message broker	Service layer	Big Data Infrastructure made available within I-BiDaaS through the Cloud Service Partner in which data are collected from different data providers
<Clause 8.1 requirement (3)> It is recommended that the CSP:BDIP supports collecting data from different CSN: DPs with different modes.	Perform data collection (7.1.3.1)	Data collection functional component (7.1.1) UM as message broker	Service layer	Big Data Infrastructure made available within I-BiDaaS through the Cloud Service Partner in which data are collected from different data providers with different modes
<Clause 8.1 requirement (6)> Data collection can optionally be performed by the CSP:BDIP in realtime.	Perform data collection (7.1.3.1)	Data collection functional component (7.1.1) UM as message broker in conjunction with Apama	Service layer	Real time for the Production process of Aluminium die-casting use case
<Clause 8.6 requirement (1)> It is required for the CSP:BDIP to manage metadata information such as creating, controlling, attributing, defining and updating.	Publish data (7.1.1.2)	Data catalogue functional component (7.3.3.3)	Operations support systems (OSS)	
<Clause 8.1 requirement (2)> It is recommended for the CSN: DP to expose data to the CSP:BDAP by publishing metadata.	Publish data (7.1.1.2)	Data catalogue functional component (7.3.3.3)	OSS	
<Clause 8.3 requirement (5)> It is recommended for the CSN: DP to expose APIs for data delivery.	Perform data storage (7.1.3.2)	Data catalogue functional component (7.3.3.3)	OSS	
<Clause 8.1 requirement (4)> It is recommended for the CSN: DP to provide a brokerage service to the CSP:BDIP for searching accessible data.	Brokerage data (7.1.1.3)	Data collection functional component (7.1.1)	Service layer	

<Clause 8.1 requirement (5)> It is recommended that the CSP:BDIP integrates data delivered by the CSC and data publicly available.	Brokerage data (7.1.1.3)	Data collection functional component (7.1.1)	Service layer	
<Clause 8.5 requirement (2)> It is recommended that the CSP:BDAP supports different tools or plug-ins with multiple styles of data visualization.	Visualize data (7.1.2.1)	Data visualization functional component (7.1.2) Qbeast AVT	Service layer	Multiple styles of visualisations are supported via the AVT including e.g., linecharts, barcharts, graph-based representations, Spatio-temporal analysis, etc
<Clause 8.5 requirement (3)> It is recommended that the CSP:BDAP supports customization of the reporting tools.	Visualize data (7.1.2.1)	Data visualization functional component (7.1.2) Qbeast AVT	Service layer	Reporting is available with multiple styles such as statistical graphics, forms, diagrams, charts by using the “results API” or the “download results” option available for all data analysis experiments
<Clause 8.5 requirement (4)> It is recommended that the CSP:BDAP supports integration of reporting tools with the CSC reporting systems.	Visualize data (7.1.2.1)	Data visualization functional component (7.1.2) Qbeast AVT	Service layer	Integration can be achieved using the “results API” offered by the orchestration service
<Clause 8.5 requirement (5)> It is recommended that the CSP:BDAP supports integration of reporting tools with the CSC operational systems.	Visualize data (7.1.2.1)	Data visualization functional component (7.1.2) Qbeast AVT	Service layer	Integration can be achieved using the “results API” offered by the orchestration service
<Clause 8.2 requirement (1)> It is required for the CSP:BDIP to support data aggregation.	Provide data integration (7.1.3.4)	Data pre-processing functional component (7.1.3) Hecuba Apama	Service layer	Data cleaning, data integration, data transformation and data extraction to improve data analysis efficiency.
<Clause 8.2 requirement (2)> It is recommended that the CSP:BDIP provides the dedicated resources for pre-processing.	Provide data pre-processing (7.1.3.3)	Data pre-processing functional component (7.1.3)	Service layer	Specifically – cleaning data which includes processing smoothing noise data, and identifying and

<Clause 8.2 requirement (3)> It is recommended that the CSP:BDIP supports unification of data collected in different formats.	Provide data pre-processing (7.1.3.3)	Data pre-processing functional component (7.1.3) Apama	Service layer	removing outliers to improve data quality; – combining and integrating data from multiple sources to remove duplicated and redundant data;
<Clause 8.2 requirement (4)> It is recommended for the CSP:BDIP to support extraction of data from unstructured data or semi-structured data into structured data.	Provide data pre-processing (7.1.3.3)	Data pre-processing functional component (7.1.3) Apama	Service layer	– transforming the data collected in different formats and types; H18 – extracting the representative features from a large number of data features for data analysis.
<Clause 8.1 requirement (2)> It is recommended for the CSN: DP to expose data to the CSP:BDAP by publishing metadata.	Publish data (7.1.1.2)	Data catalogue functional component (7.3.3.3)	Service layer	
<Clause 8.3 requirement (5)> It is recommended for the CSN: DP to expose APIs for data delivery.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5)	Service layer	
<Clause 8.4 requirement (1)> It is required for the CSP:BDAP to support analysis of various data types and formats.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs Hecuba Qbeast Apama	Service layer	UNSPMF's advanced ML submodule can perform data analysis for the various use cases and the different data types and formats supported by the I-BiDaaS platform
<Clause 8.4 requirement (2)> It is required for the CSP:BDAP to support batch processing.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs Hecuba Qbeast	Service layer	UNSPMS's advanced ML submodule, in conjunction with BSC's COMPSs can perform batch processing.
<Clause 8.4 requirement (3)> It is required for the CSP:BDAP to support association analysis.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4)	Service layer	Tasks related to association analysis can be supported by the Advanced ML submodule
<Clause 8.4 requirement (4)> It is required for the CSP:BDAP to support different data analysis algorithms.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs	Service layer	Various use cases with different analyses and algorithms
<Clause 8.4 requirement (5)>	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4)	Service layer	

It is recommended that the CSP:BDAP supports customization of analytical applications.		COMPSs		
<Clause 8.4 requirement (6)> It is recommended for the CSP:BDAP to support user defined algorithms.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs	Service layer	
<Clause 8.4 requirement (7)> It is recommended for the CSP:BDAP to support data processing in distributed computing environments.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs Hecuba Qbeast	Service layer	
<Clause 8.4 requirement (9)> It is recommended that the CSP:BDAP supports data classification in parallel.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs Hecuba Qbeast	Service layer	
<Clause 8.4 requirement (10)> It is recommended that the CSP:BDAP provides different analytical applications.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs	Service layer	Different analytical applications are supported by different implementations of UNSPMF's algorithms, developed both as general purpose and for specific industrial use cases. E.g. - quality improvement - predictive maintenance - cost reduction
<Clause 8.4 requirement (11)> It is recommended that the CSP:BDAP supports customization of analytical applications.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs	Service layer	
<Clause 8.4 requirement (12)> It is recommended for the CSP:BDAP to support real-time analysis of streaming data.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) Hecuba Qbeast Apama streaming analytics	Service layer	Production process of Aluminium die-casting

<Clause 8.4 requirement (13)> It is recommended for the CSP:BDAP to support user behavior analysis.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4)	Service layer	Customization of analysis supports the customization of detail data analysis methods according to a data provider's specific requirements.
<Clause 8.4 requirement (14)> The CSP:BDAP can optionally perform analysis of different data types and formats in realtime.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) Apama streaming analytics	Service layer	
<Clause 8.6 requirement (2)> It is required for the CSP:BDIP to track a data history which contains source of data and data processing method.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4)	Service layer	
<Clause 8.3 requirement (1)> It is required for the CSP:BDIP to support different data types with sufficient storage space, elastic storage capacity, and efficient control methods.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5)	Service layer	
<Clause 8.3 requirement (2)> It is required for the CSP:BDIP to support storage for different data formats and data models.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5) COMPSs Hecuba Qbeast	Service layer	
<Clause 8.3 requirement (4)> It is recommended that the CSP:BDIP provides different types of databases.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5) Hecuba	Service layer	
<Clause 8.4 requirement (8)> It is recommended for the CSP:BDAP to support data indexing.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5) Qbeast	Service layer	
<Clause 8.4 requirement (7)> It is recommended for the CSP:BDAP to support data processing in distributed computing environments.	Manage data provenance (7.1.3.6)	Data storage functional component (7.1.5) COMPSs Hecuba Qbeast Apama GPU integration	Service layer	Advanced ML submodule supports distributed computing environments based on the BSC's COMPSs framework

<Clause 8.5 requirement (6)> It is recommended that the CSP:BDAP supports composed services which could combine two or more big data services to the CSC: BDSU.	Use big data service (7.1.4.1)	Third-party service integration functional component (7.3.1.1)	Integration	
<Clause 8.4 requirement (6)> It is recommended for the CSP:BDAP to support user defined algorithms.	Analyze data (7.1.2.2)	Data analysis functional component (7.1.4) COMPSs Apama streaming analytics	Service layer	UNSPMF's algorithms can be user-defined in the I-BiDaaS expert mode of the platform, by coding a new implementation based on available templates; or user can tune hyperparameters, or select a pre-defined algorithm in the self-service mode
<Clause 8.7 requirement (2)> It is required for the CSP:BDIP to support data protection.	Manage data protection (7.1.3.5)	Security and privacy management functional component (7.3.2.1)	Security systems	
<Clause 8.7 requirement (5)> It is recommended that the CSP:BDIP supports redundancy mechanism and transaction logging.	Use big data service (7.1.4.1)	Cross-cutting aspect (8.1) Hecuba	Multiple layers for cross-cutting aspect	
<Clause 8.7 requirement (1)> It is required for the CSP:BDIP to protect data collection, data storage, data transmission, and data processing with security mechanisms.	Manage data protection (7.1.3.5)	Security and privacy management functional component (7.3.2.1)	Security systems	
<Clause 8.3 requirement (6)> It is recommended that the CSP:BDIP fulfils storage and database performance demands.	Perform data storage (7.1.3.2)	Cross-cutting aspect (8.2) Hecuba Qbeast	Multiple layers for cross-cutting aspect	
<Clause 8.6 requirement (3)> It is required for the CSP:BDAP to support distributed cluster monitoring tools to monitor the health and	—	Distributed processing functional component (7.2.1)	Resource layer	



status of computing clusters.				
<Clause 8.6 requirement (5)> It is recommended for the CSP:BDIP to support network resource monitoring.	–	Distributed processing functional component (7.2.1)	Resource layer	
<Clause 8.3 requirement (3)> It is required that the CSP:BDIP provides flexible licensing policy for the database.	Use big data service (7.1.4.1)	Data policy management functional component (7.3.3.2)	OSS	
<Clause 8.3 requirement (7)> It is recommended that the CSP:BDIP supports data retention policy covering data retention period before its destruction after termination of a contract, to protect the big data service customer from losing private data through an accidental lapse of the contract.	Manage data protection (7.1.3.5)	Data policy management functional component (7.3.3.2)	OSS	
<Clause 8.7 requirement (4)> It is recommended that the CSP supports implementing the CSC's data protection and security policies over data and analytical results.	Manage data protection (7.1.3.5)	Data policy management functional component (7.3.3.2)	OSS	
<Clause 8.7 requirement (3)> It is required that the CSP deletes CSC related data and analytical results according to the lifetime defined by the CSC or on the CSC's demand.	Manage data protection (7.1.3.5)	Data policy management functional component (7.3.3.2)	OSS	
<Clause 8.6 requirement (6)> It is recommended for the CSP:BDIP to support management of data life-cycle operations.	–	Data life-cycle monitoring functional component (7.3.3.1)	OSS	

<Clause 8.6 requirement (4)> It is required for the CSP:BDIP to support data preservation policy management rules.	–	Data life-cycle monitoring functional component (7.3.3.1)	OSS	
<Clause 8.1 requirement (4)> It is recommended for the CSN: DP to provide a brokerage service to the CSP:BDIP for searching accessible data.	Brokerage data (7.1.1.3)	Data collection functional component (7.1.1)	Service layer	
<Clause 8.3 requirement (1)> It is required for the CSP:BDIP to support different data types with sufficient storage space, elastic storage capacity, and efficient control methods.	Perform data storage (7.1.3.2)	Resource orchestration functional component for big data (7.3.3.4)	OSS	
<Clause 8.3 requirement (6)> It is recommended that the CSP:BDIP fulfils storage and database performance demands.	Perform data storage (7.1.3.2)	Data storage functional component (7.1.5) Hecuba Qbeast	Service layer	
<Clause 8.3 requirement (3)> It is required that the CSP:BDIP provides flexible licensing policy for the databases.	Perform data storage (7.1.3.2)	Data policy management functional component (7.3.3.2)	OSS	
<Clause 8.4 requirement (5)> It is required that the CSP:BDAP provides a flexible licensing policy for the analytical applications.	Analyze data (7.1.2.2)	Data policy management functional component (7.3.3.2) COMPSs	OSS	
<Clause 8.5 requirement (1)> It is required that the CSP:BDAP provides a flexible licensing policy for the reporting tool.	Publish data (7.1.1.2)	Data policy management functional component (7.3.3.2)	OSS	