Horizon 2020 Program (2014-2020)

Big data PPP

Research addressing main technology challenges of the data economy

I-BiDaaS

Industrial-Driven Big Data as a Self-Service Solution

Industrial-Driven Big Data as a Self-Service Solution

D6.2: Experiments implementation – initial version[†]

Abstract: This is a report on the progress of: the operational experiments' specification; the preparation of the datasets to be exploited in each experiment; and the indicators to be measured in order to validate both the operational (business) and technical performance of the I-BiDaaS platform, according to the I-BiDaaS experimental protocol. Furthermore, it provides an overview of the first completed I-BiDaaS prototype and the activities aiming towards the involvement of external stakeholders in the experimentation process.

Contractual Date of Delivery	30/06/2019
Actual Date of Delivery	30/06/2019
Deliverable Security Class	Public
Editor	*Evangelia Kavakli (UNIMAN),* *Rizos Sakellariou (UNIMAN)*
Contributors	All *I-BiDaaS* partners
Quality Assurance	*Giuseppe Danilo Spennacchio (CRF),* *Enric Pages (ATOS)* *Kostas Lampropoulos (FORTH)*

[†] *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. *Enric Pages, (ATOS)*
3. *Kostas Lamproulos (FORTH)*

Revisions

Version	Date	By	Overview
1.4	26/06/2019	E. Kavakli	Final. Revised after CRF's second round of review. Incorporated latest updates on experiments' progress.
1.3	14/06/2019	E. Kavakli	Second draft. Incorporate reviewers comments
1.2	30/05/2019	E. Kavakli	First draft. Include input from technology providers and use case providers
1.1	08/05/2019	E. Kavakli	Comments on the ToC.
1.0	22/04/2019	E. Kavakli	ToC.

Table of Contents

LIST OF TABLES 5

LIST OF FIGURES 6

LIST OF ABBREVIATIONS 7

EXECUTIVE SUMMARY 8

1 INTRODUCTION 9

2 EXPERIMENTAL PROTOCOL ALIGNMENT 10

3 IMPLEMENTATION AND OPERATION OF REAL-LIFE INDUSTRIAL EXPERIMENTS 13

3.1 OVERVIEW 13
3.2 DATASETS PREPARATION 15
3.2.1 Synthetic Datasets preparation *15*
3.2.2 Real Datasets preparation *16*
3.2.3 TID Datasets progress *16*
3.2.4 CAIXA Datasets progress *16*
3.2.5 CRF Datasets progress *20*
3.3 OPERATIONAL EXPERIMENTS DEFINITION 21
3.3.1 Experiment workflow *21*
3.3.2 Telecommunication experiments *22*
3.3.3 Banking experiments *23*
3.3.4 Manufacturing experiments *26*
3.3.5 Cross-sectoral experiment definition *27*

4 EXPERIMENTAL EVALUATION 29

4.1 OVERVIEW 29
4.2 THE FIRST RELEASE OF THE INTEGRATED I-BIDAAS SOLUTION 29
4.2.1 Wireframes of the I-BiDaaS solution for the functionality of standard/default experiment *33*
4.3 EXPERIMENTS VERIFICATION AND VALIDATION 36
4.4 QUANTITATIVE EVALUATION 36
4.5 QUALITATIVE EVALUATION 42
4.5.1 Telecommunication experiments *42*
4.5.2 Banking experiments *43*
4.5.3 Manufacturing experiments *45*
4.5.4 Common experimental indicators *46*

5 IMPACT ANALYSIS 48

5.1 PROGRESS REPORT 48
5.2 INVOLVING EXTERNAL STAKEHOLDERS IN THE EVALUATION PROCESS 50
5.2.1 I-BiDaaS external advisory board – the introductory telcos *50*
5.2.2 I-BiDaaS external advisory board – the first physical meeting *51*
5.2.3 1st I-BiDaaS Info Day-Workshop in Novi Sad *52*
5.2.4 CRF's hackathon at Campus Melfi *52*
5.2.5 Hackathon at TID *54*

6 CONCLUSION 55

7 REFERENCES 56

ANNEX A QUANTITATIVE EVALUATION QUESTIONNAIRE 57

ANNEX B QUALITATIVE EVALUATION QUESTIONNAIRE 59

List of Tables

Table 1: Overview of use cases and associated Data Sets 14
Table 2: Analysis of relationships through IP address dataset structure 16
Table 3: Advanced analysis of bank transfer payment in financial terminal dataset structure 17
Table 4: Datasets of the I-BiDaaS integrated solution 30
Table 5: Hardware specification 31
Table 6: Component evaluation 37
Table 7: Integrated solution evaluation 39
Table 8: Classification of applicable Big Data benchmarking tools 40
Table 10: Progress with regards to I-BiDaaS KPIs 49
Table 11: Feedback received from the EAB members 51

List of Figures

Figure 1: I-BiDaaS experimental process.......... 10
Figure 2: Dice comparison of the bloom filtering results (with regards to the word 'MALADETA'). 19
Figure 3: Generic experimental workflow 21
Figure 4: Revised data flow.......... 30
Figure 5: The batch processing flow.......... 32
Figure 6: The streaming processing flow.......... 33
Figure 7: Standard/default experiment for use case 1.......... 34
Figure 8: Standard/default experiment for use case 2.......... 34
Figure 9: Run standard experiment.......... 35
Figure 10: Create/Update custom project.......... 35
Figure 11: Run custom experiment.......... 36

List of Abbreviations

AMQP	Advanced Message Queueing Protocol
API	Application Programming Interface
ASR	Automatic Speech Recognition
BDV	Big Data Value
BDVA	Big Data Value Association
CEO	Chief Executive Officer
CPU	Central Processing Unit
CSI	Customer Satisfaction Index
CSP	Constraint Satisfaction Problem
CSV	Comma-separated Values
CTM	Time Marked Conversation
DB	Data Base
EAB	External Advisory Board
(GAM)	Generalized Additive Model
GDPR	General Data Protection Regulation
JPH	Job Per Hour
JMS	Java Messaging Service
IOPs	Input/output operations per second
KPI	Key Performance Indicator
MB	Megabyte
MLP	Multi Layer Perceptron
MPI	Message Passing Interface
MQTT	Message Queueing Telemetry Transport
MVP	Minimum Viable Product
NLP	Natural Language Processing
OEE	Overall Equipment Effectiveness
PPP	Public-Private Partnership
RTTM	Rich Transcription Time Marked
SMEs	Small and Medium-Sized enterprises
SQL	Structured Query Language
SRIA	Strategic Research and Innovation Agenda
TDF	Test Data Fabrication, formerly known as DFP
TN	True Negative
TP	True Positive
UM	Universal Messaging
VM	Virtual Machine
WP	Work Package

Executive Summary

This deliverable is the baseline of the I-BiDaaS experimentation phase (M18-M30). In particular, it reports on the progress of the experiments' specification as well as on the preparation of the datasets to be exploited in each experiment. In addition, it provides a detailed list of indicators to be measured and benchmarks to be used during the experiments in order to validate both the operational (business) and technical performance of the I-BiDaaS platform. Special emphasis is placed on ensuring the alignment of the operational experiments with the business objectives of the industrial partners expressed in the context of the defined use cases. Finally, it provides an initial impact analysis report with respect to the expected project innovation and achievements, as well as an overview of the current and prospective activities towards the involvement of external stakeholders, (both domain experts and prospective users), in the evaluation process. The operational experiments' specification reported in this deliverable will be further refined and adjusted during the experimentation phase to reflect data availability and platform functionality.

1 Introduction

This deliverable presents the initial results of "*WP6. Real – life industrial and operational experiments*". In particular, it describes the experimental protocol alignment process aiming to ensure a smooth and adequate running of the experiments according to the experimental protocol. Furthermore, it reports on the progress of the preparation of the datasets to be exploited during the experimental process. In addition, it provides an updated specification of the different experiments that will be performed in the context of the real industrial use case scenarios, defined during the project baseline phase. Finally, it provides structured feedback from the technology owners, to the development of the first complete I-BiDaaS prototype and the updated list of indicators and benchmarks to be used for validating the technical and operational performance of the first I-BiDaaS platform, during the experimentation phase of the project (M18-M30).

The deliverable also provides an initial impact analysis with respect to the expected project level innovation and achievements, as well as a description of the current and prospective activities aiming to the involvement of external stakeholders (both domain experts and members of the wider user community) in the evaluation process.

This deliverable synthesizes the initial results of all WP6 activities and as such, there is a close interrelation between this deliverable and all the tasks in the current work package. Furthermore, there is a close dependency between this deliverable and all other work packages. In particular, the results of WP1 and specifically the results of T1.1 and T1.4 are used as the basis for the alignment of the industrial real-life experiments (in T6.1), whilst the I-BiDaaS prototype releases resulting from WP5, and especially tasks (T5.3 and T5.4) will be used in the implementation and operational of the experiments (T6.2). Furthermore, the results of impact analysis of T6.3 will provide input towards potential exploitation activities of the project in the context of WP7 (T7.3).

The deliverable is directly related to the achievement of the I-BiDaaS objective, *'to develop, validate, demonstrate, and support, a complete and solid Big Data solution that can be easily configured and adapted by practitioners'*. In addition, the specifications of the different experiments are aligned with the specific objectives of the I-BiDaaS project towards *'the development of solutions applicable in real-world settings that demonstrate a significant increase of speed of data throughput and access'*, and the *'demonstration of how data fabrication can help companies in their Big Data experimentation and testing'*.

The remainder of this report is structured as follows. Section 2, describes the experimental protocol alignment process performed in order to refine and if necessary revise the outcomes of the initial experiment's definition phase according to the I-BiDaaS experimental protocol and to fine-tune the details of the industrial experiments to assure that the designed experiments will validate both business and technical requirements. Section 3 provides an update on the current implementation and operation of the real-life industrial experiments, reporting on the progress of the dataset preparation as well as the experiments' specification. The progress regarding the quantitative and qualitative evaluation of the I-BiDaaS solution at component, platform and experiment level is reported in section 4. Section 5, discusses the impact so far and provides an overview of current and prospective external stakeholders involvement activities. Finally, section 6 concludes with a summary of the current achievements of WP6 and pointers to future work.

2 Experimental protocol alignment

The I-BiDaaS experimental protocol (reported in D1.3) aims at evaluating and validating the I-BiDaaS architecture and its implementation in the specific project use cases. According to this protocol the experimental process (depicted in Figure 1) consists of:

- Two *'definition'* phases namely those of *scoping* and *planning* of the experiments in terms of defining the context of the experiments (goals, tasks, participants) as well as identifying the business and technology indicators to be measured. In addition, applicable technical benchmarks as well as business improvements sought are identified in these initial phases.
- Two *'operational'* phases namely those of *operation* and *analysis* which focus on the implementation of the real-life experiments and the quantitative and qualitative evaluation of the I-BiDaaS platform at component, system and experiment level.

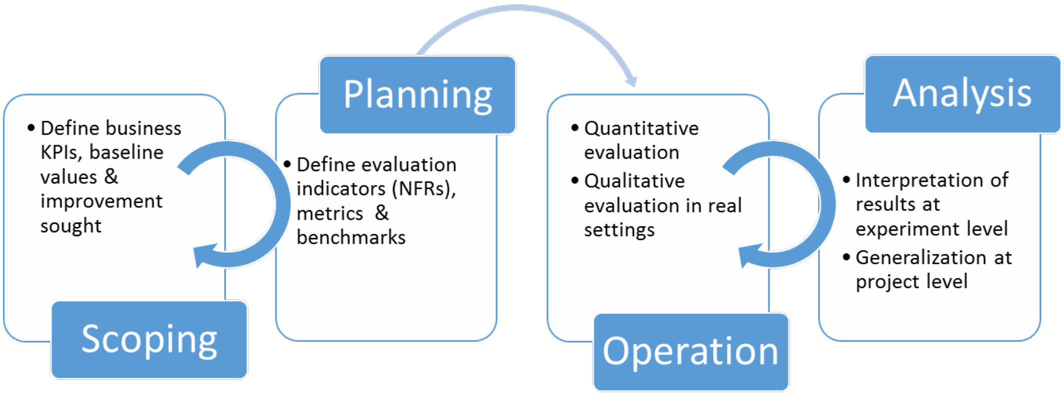

Scoping
• Define business KPIs, baseline values & improvement sought
Planning
• Define evaluation indicators (NFRs), metrics & benchmarks
Operation
• Quantitative evaluation
• Qualitative evaluation in real settings
Analysis
• Interpretation of results at experiment level
• Generalization at project level

Figure 1: I-BiDaaS experimental process.

The experimental validation, focuses mainly on I-BiDaaS quality aspects (non-functional requirements) and is a separate though complementary process to the verification (testing) process performed at component and system level, according to the system functional requirements (described in D5.2, section 2.3).

A main characteristic of the experimental process is that it considers both technical and business requirements. In particular, the definition of each experiment considers both aspects in an integrated way, aiming to validate not only the performance of the I-BiDaaS solution, but also its alignment with the needs of the industrial users. Thus, it requires the collaboration of both technology and business stakeholders in all phases.

Another important aspect is that it is an iterative and incremental process, whereby the definition of the experiments is iteratively refined to increased level of certainty in order to reflect (a) revisions of the industrial use cases definitions and associated requirements and (b) progress in the design of the I-BiDaaS platform and associated technology characteristics. This is referred to as '*experimental alignment*' and aims to ensure that the designed experiments reflect the actual business and technical requirements.

For example, let us consider the *'Analysis of relationships through IP address'* banking use case. Through this use case CAIXA wishes to improve the process of testing novel tools and in particular to speed up the implementation of new Big Data Analytics solutions. So the initial KPI defined during the scoping phase has been *'time to implement new technologies'.* Further analysis of the use case revealed that one of the main factors delaying the implementation of new technologies is the bureaucracy due to the data privacy concerns, which results in delays in providing access to the data set to external technology providers.

Therefore, the initial KPI is refined to better reflect the business requirements, as *'time of grating permits for data access at the start of a project with an external provider'*. To this end, the I-BiDaaS solution considers the use of synthetic data through IBM's test data fabrication component. Using synthetic data CAIXA aims to reduce time for granting permits to data access at the start of a project with an external provider. This of course, depends on the ability of fabricated data to provide the same or additional insights compared to the real data. Therefore the goal of the associated experiment becomes *'to validate the quality and efficiency of the use of synthetic data for analysis'*, thus leading to the identification of applicable performance indicators at the application level (e.g., *validity of generated data model, time for generating a volume of synthetic data that can provide a valid model*).

The above example demonstrates that the *traceability* among use case KPIs, experiments' goals and experimental indicators is assured.

To assist this alignment process, two questionnaires (see Annexes A and B) were developed, based on the initial results of the scoping and planning phases reported in D1.3, as well as state of the art in big data benchmarking:

- A *quantitative evaluation* questionnaire, focusing on the evaluation of the technical quality of the I-BiDaaS platform in parts and as a whole using appropriate indicators and associated metrics. The main stakeholders using this questionnaire were the technology providers.
- A *qualitative evaluation* questionnaire, focusing on the experimental evaluation of the I-BiDaaS platform in real business settings and against user requirements (defined during the planning phase). The main stakeholders using this questionnaire were industrial users.

The aim of these questionnaires has been to reach an agreed set of indicators against which validation of the 'success' will be measured. Such indicators may reflect technology features at component and platform level (e.g., operational performance), or business key performance indicators (e.g., customer satisfaction, time efficiency). The former are use case independent, whilst the latter reflect the specific needs expressed in each use case. Alignment of both types of indicators is a main objective of the experimental definition phase.

For each indicator, a set of quantifiable metrics is also defined, whose measurement relates to the achievement (or not), of a specific indicator. For example, in the case of operational performance, relevant metrics include execution time and throughput. In the case of customer satisfaction a relevant metric might be the 'low customer satisfaction index (CSI)'. For the metrics related to technology indicators, available big data benchmarks can also be used. An initial investigation of applicable big data benchmarks has been performed during the I-BiDaaS baseline phase and reported in D1.3. This has been further informed from recent research in the area, reported in the results of ongoing projects, such as the classification of big data benchmarks proposed by the European DataBench[1] project, in which some of the most well-known benchmarking tools are classified according to the benchmark categories (Micro and Application benchmarks), type, domain, data type and metrics measured.

The completion of these questionnaires is an ongoing process which also follows an iterative approach leading to several refinements performed during online workshop sessions as well as

[1] https://www.databench.eu/

during online interviews with the use case providers. An overview of the data collected so far is reported in sections 4.4 and 4.5.

The results of this process will also form the baseline for the preparation of the checklists to be used for the recording of the experimental results during the operation phase. Once again, this will be an iterative process, following the I-BiDaaS prototype releases.

3 Implementation and operation of real-life industrial experiments

3.1 Overview

I-BiDaaS real-life industrial experiments are defined on the basis of the real use case scenarios within the three industrial sectors where the project's industrial partners are involved. Each experiment will exploit certain dataset(s) defined during the process set up phase and reported in D1.3 and D2.1. In addition complementary experiments can be defined in the same use case, for example utilizing different datasets (synthetic / real data) or processing type (batch / streaming).

The first step in the implementation and operation of the real-life industrial experiments is the preparation of the datasets using the methods developed in WP2 to make it ready for use within the I-BiDaaS solution. This in certain cases required the revision of the datasets specification.

In particular, CAIXA planned at the beginning of the project to generate synthetic data using IBM Test Data Fabrication (TDF) for all the different use cases. However, after the generation of the synthetic dataset for the first use case and the data analysis process, it has been decided to provide real tokenized data for the rest of the use cases. The data preparation process for the data serialization and encryption implies much more effort but the expected benefits of using real data are much higher.

Therefore, it had some implications with regards to CAIXA initial data generation plan. On the one hand, it has been decided to focus on three use cases and skip the social graph case initially defined. This change was proposed by CAIXA considering the increment of the amount of work, especially in the data preparation and in-depth analysis of the data. Working with real data implied a time-consuming involvement of CAIXA in these processes and the extraction of a model for each use case that could be used in day-by-day business operation, and not only for testing purposes as previously planned.

On the other hand, some difficulties on accessing relevant data were found in the process of collecting real data from the external providers control access use case. Therefore, this use case is still focused on the analysis of connectivity patterns but taking into account the customer online banking accesses instead of the third-party provider ones to CAIXA platforms.

Nevertheless, the generation of data for the three resulting CAIXA use cases is underway and it is detailed in section 3.2.4.

Table 1: Overview of use cases and associated Data Sets

No.	Use Case	I-BiDaaS dataset	Data set Preparation Status	Data Provider
1	Accurate location prediction with high traffic and visibility	Anonymized TID mobility data Synthetic TID mobility data	Not started - 0% Generated - 100%	TID
2	Optimization of placement of telecommunication equipment	Anonymized TID mobility data Synthetic TID mobility data	Not started - 0% Generated - 100%	TID
3	Employment of bots in call center	Synthetic call center data	In progress - 30%	TID
4	Enhance control over online banking access	Online banking control tokenized data	NEW Not started - 0%	CAIXA
5	Advanced analysis of bank transfer payment in financial terminal	Bank transfer tokenized data	In progress - 90%	CAIXA
6	Analysis of relationships through IP address	Synthetic IP address data Tokenized IP address data	Generated - 100% In progress - 70%	CAIXA
7	Building of a social graph	Synthetic Social graph data	DROPPED	CAIXA
8	Maintenance and monitoring of production assets	Anonymized SCADA data Synthetic SCADA data Anonymized MES data Synthetic MES data	In progress - 70 % Not started - 0% In progress - 15% Not started - 0%	CRF
9	Production process of aluminium casting	Anonymized Aluminium Casting data Synthetic Aluminium Casting data	Generated - 100% Generated - 100%	CRF

Table 1 provides an overview of the datasets to be exploited and their preparation status, further detailed in section 3.2.

Following the dataset preparation, the operational experiments and trials will be carried out using the I-BiDaaS solution, according to the experiment's definition in terms of:

- ***Experiment's goals***: they define the purpose of the experiment. Although the experimental goals are related to the use case objectives they are not necessarily identical. Potentially, multiple experiments might be defined in the context of the same use case. These experiments are complementary in the sense that the corresponding experimental goals mutually support the achievement of use case goals.
- ***Experimental questions***: they define what the experiments specifically aim to measure. These questions will guide the definition of the *experimental indicators and metrics* which define the type of data being recorded during the experiment in order to answer the experimental questions.

- ***Experiment's workflow***: it defines the type and order of activities (workflow) that will be involved in each experiment.
- ***Experimental subjects***: the type of users (roles) to be involved in the experiment.

Section 3.3, report~~s~~ on the current status of the experiments' definition at M18. It should be noted that this is an ongoing process and the experiment's definition will be further refined during the experimentation phase of the project.

3.2 Datasets preparation

3.2.1 Synthetic Datasets preparation

WP2 provides a specification for the I-BiDaaS platform's heterogeneous datasets of the finance, manufacturing, and telecommunications targeted domains. WP2 is based on WP1 achievements which provides aggregated and ingested data sets, the processes and activities related to the organization and integration of the data collected for the various use cases and the definition of what data needs to be complemented synthetically.

D2.1 provides, for each use case, an overview containing brief explanation of the objectives and goals followed by a description of the used datasets nature and formats and a section containing a concrete definition of the dataset actual structure.

Using these descriptions, definitions, formats and data relationships, we were able to create three data projects, one for each of the three synthetic datasets: CAIXA IP address, TID mobility data and CRF manufacturing.

The process of generating each of the datasets included the creation of the data project and the fabrication itself using that data project. The task of creating a data project included configuring the proper resources and importing them into the project, "translating" the provided requirements into formal constraints (rules) that will be used by the CSP solver for data generation and finally creating fabrication and configuration for the desired dataset involving the selected resources and constraints.

A few examples for the definitions that had to be "translated" into rules were:

- Generate some number (any number) of different users.
- Generate some number of random IPv4 addresses per user.
- Generate some minimal number of rows with the same IPv4 address but with different connection time.
- Generate ~5% duplicates for the various IPv4 addresses.
- Generate process parameter set of values using the Max/Min/Median/Mean/Variance etc. An example of a process parameter could be signal, pressure, heat etc.
- Generate a unique ID (user is, sector is, event is etc).
- Generate some number of events per sectors or per users using a given distribution (could also be discrete).

The actual fabrication required installation and configuration of PostgreSQL DB on the dedicated VM including the creation of the relevant schemas and tables as defined in D2.1.Once everything was setup properly, the fabrication process could be started. The data was also generated into SQLite DB. In both cases the generated data was exported into CSV files.

Currently the CAIXA IP address dataset contains 500K records, the TID Mobility data dataset contains 700K records and the CRF manufacturing dataset contains 1.2KK records.

3.2.2 Real Datasets preparation

Real data set preparation involves the following processes. The first activity is the identification of all the data sources (business process data tables) that relate to the dataset description. Second, the real data provided are anonymized, masked, tokenized and manipulated in various ways, depending on the type of the data field, to avoid privacy regulations violations as well as to avoid leakage of certain sensitive business details. The objective is to lose as less information as possible and enable the data analysis without decrypting the data.

Currently, the preparation of the CAIXA IP address tokenised data and Bank Transfer tokenised data are in progress; the CRF anonymised Aluminium Casting dataset has been generated. More details are provided in the following sections 3.2.3 - 3.2.5.

3.2.3 TID Datasets progress

So far, the generation of the synthetic Mobility dataset (to be exploited in the context of the "Accurate location prediction with high traffic and visibility" and "Optimization of placement of telecommunication equipment" use cases), according to the specification reported in D2.1 has been completed using the IBM TDF tool. The latest version of the data contains 700k entries.

The process of providing in-house access to two real mobility datasets is in progress. So is the provision of a real dataset related to the "Employment of bots in call center" use case.

3.2.4 CAIXA Datasets progress

The generation of a synthetic dataset for "Analysis of relationships through IP address" use case is already done by means of IBM TDF tool and the generated dataset was used to test first version of I-BiDaaS Minimum Viable Product (MVP). The preparation and tokenization of real data for this use case is also in progress (process at 70%) in order to compare the results of the real and synthetized data. The data of this dataset is stored into a single table as shown in *Table 2*.

Table 2: Analysis of relationships through IP address dataset structure

Attribute	Description	Format	Example
FK_NUMPERSO	Identifier of the Person.	NUMBER	34523454
PK_ANYOMESDIA	Date (YYYYMMDD) of the connection of the user.	NUMBER	20180823
IP_TERMINAL	IP Address of the connection of the user.	VARCHAR2	10.8.2.22
FK_COD_OPERACION	Code of business operation done by the user.	VARCHAR2	CA.OFI.ContrataActivoConSol.MenuConSolicitud
PK_COD_ESTADO_OP	Code of the status of the operation done by the user.	VARCHAR2	KO

An additional dataset related to this use case was generated for testing the streaming data analytics features of the MVP, using the generated relationship from users to validate bank transfers executed from different channels of the bank (online banking, bank offices, etc.). This process is further detailed in I-BiDaaS deliverable D5.2.

Furthermore, the collection, preparation and anonymization of the data for "Advanced analysis of bank transfer payment in financial terminal" use case is almost finished at the time of this report (only missing the validation of the tokenized data extraction from CAIXA premises). The final structure of the table for this dataset is shown in Table 3.

Table 3: Advanced analysis of bank transfer payment in financial terminal dataset structure

Attribute	Description	Format	Example
EFK_CENTRO_AP	Bank office where the transfer is done.	NUMBER	1
EFK_CONTRATO_PPAL_OPE	Sender account contract number.	NUMBER	201757974702
EFK_EMPLEADO	Identifier of the employee.	VARCHAR2	K9232951
EFK_EMPLEADO_AP	Identifier of the employee who opened the session in the terminal.	VARCHAR2	K9232951
EFK_EMPLEADO_AUT	Identifier of the employee who authorize the operation.	VARCHAR2	K3371752
EFK_NUMPERSO_PRINCIPAL	Client identifier.	NUMBER	1501630
EFK_IMPORTE_PRINCIPAL	Amount of money of the bank transfer	NUMBER	349101
EIP_TERMINAL	IP address of the terminal.	VARCHAR2	00.85.80.409
EPK_COD_TIPO_REGISTRO	Operation code identifier	VARCHAR2	657;657;657;657;657;657;657; …
EPK_OFICINA_EJEC_CAUSA	Bank office who executes the operation.	NUMBER	1
EPK_TERMINAL	Terminal identifier number.	VARCHAR2	1FRYLTOT
ETXT_AUTORIZ	Additional text of the authorization. It explains the authorization, if needed.	VARCHAR2	2;3;9;10;24;37;39;41;48; …
ECONTRATO_DESTINO	Receiver account contract number.	NUMBER	20176850664981
EAPELLIDO1_EMPLEADO	1st surname of the sender.	VARCHAR2	61;102;126;176;178;194;201; …
EAPELLIDO2_EMPLEADO	2nd surname of the sender.	VARCHAR2	17;22;28;37;38;40;49;91;93; …
ENIL_EMPLEADO	National identifier number of the employee.	VARCHAR2	09009273U
EFECHA_INI_FUNCION_EMP	Date which the employee starts working in the bank.	NUMBER	66900526
EFK_PROVINCIA	Province code identifier.	NUMBER	382
EFK_PAIS	Country code identifier.	NUMBER	35
ECOD_BLOQUE_O_EDIFICIO	Building code identifier.	VARCHAR2	['0']
ECOD_TIPO_VIA	Sender physical address street type.	VARCHAR2	-1
EFK_CENTRO_GESTOR	Identifier of the usual bank office of the sender.	NUMBER	1
EFK_CENTRO_RELACIONADO_1	Additional bank office of the sender 1.	NUMBER	01
EFK_CENTRO_RELACIONADO_2	Additional bank office of the sender 2.	NUMBER	-1

EFK_CENTRO_RELACIONADO_3	Additional bank office of the sender 3.	NUMBER	-1
EFK_CENTRO_RELACIONADO_4	Additional bank office of the sender 4.	NUMBER	-1
EFK_CODIGO_POSTAL	Post code of the sender.	NUMBER	31683
EFK_EMPLEADO_GESTOR	Identifier of the usual employee assigned to the sender.	NUMBER	75565
EFK_EMPLEADO_GESTOR_2	Identifier of the deputy employee assigned to the sender.	NUMBER	-1
EFK_PAIS_RESIDENCIA	Residence country of the sender.	NUMBER	35
EFK_PROVINCIA_DOMICILIO	Residence province of the sender.	NUMBER	17256
ENOMBRE_LOCALIDAD_DOMICIL	Residence town/city of the sender.	VARCHAR2	Dzpmrcjr
ENOMBRE_PERSONA_PFISICA	Name of the sender	VARCHAR2	Ykshturywsow
ENOMBRE_VIA	Physical address of the sender (street name)	VARCHAR2	29;80;92;99;162;194;205;213;…
ENUM_DESDE	Physical address of the sender (initial street number).	NUMBER	6
ENUM_EDAD	Age of the sender	NUMBER	2992554
ENUM_EDIFICIO	Physical address of the sender (building number).	NUMBER	-1
ENUM_ESCALERA	Physical address of the sender (stairs number).	NUMBER	-1
ENUM_HASTA	Physical address of the sender (end street number).	NUMBER	-1
ENUM_PISO	Physical address of the sender (floor number).	NUMBER	-1
ENUM_PUERTA	Physical address of the sender (door number).	NUMBER	-1
EOTROS_DATOS_DIRECCION	Physical address of the sender (additional data).	VARCHAR2	-1
EPK_EMPRESA_GRUPO	Bank branch identifier.	NUMBER	1
EPRIMER_APELLIDO_PFISICA	1st surname of the sender.	VARCHAR2	5;7;34;48;50;62;69;78;100;119;…
ERAZON_SOCIAL	Name of the sender (if it is an enterprise)	NUMBER	-1
ESEGUNDO_APELLIDO_PFISICA	2nd surname of the sender	VARCHAR2	6;11;13;16;21;108;113;126;133; …
EPK_TIPREL	Identifier with the type of relationship with the bank.	NUMBER	7

For generating this table, three processes were followed. First the identification of all the business process tables that relates to the events produced by the employee for executing the bank transfer. After that, this information was joined with the information of other tables adding information of the bank office, employee, sender, receiver of the bank transfer, etc. Once we have the enhanced table with all the contextual information from the bank transfer available, the tokenization process is executed.

For this process three different types of encryptions are used depending on the type of the field in order to lose the less information possible and enabling the data analysis outside of CAIXA premises without decrypting the data (that is why the term tokenized is used, so the data analyst outside CAIXA environment will never be able to decrypt the data).

After an in-depth analysis of the different types of data, the following three types of encryptions were used:

- **Format Preserving Encryption**[2]. This method helps us to encrypt texts and integers in such a way that the length / format are preserved.
    - Example: Encrypt the FK_EMPLEADO (employee) field. U0133412 → K9232951.
- **Order Preserving Encryption**[3]. This method helps us to encrypt in such a way that if I have two integers $N < M$, after encrypting the order is maintained, $N' < M'$.
    - Example: Encryption of the ENUM_EDAD (age) field.
- **Bloom filters**[4,5]. This method helps us to encrypt in such a way that if we have two similar texts, after the encryption, they remain similar. It has been used for encrypting text.
    - Example: ENOMBRE_VIA (street name) = 'MALADETA'. Figure 2 shows a metric (Dice coefficient) that measures the similarity between two VIAS [e.g. It says ('MALADETA', 'CALLE MALADETA') ~ 0.8]. This helps identify similarities between free text fields once encrypted. A priori, this method is not reversible.

$$\boxed{\text{Dice Coefficient} = \frac{2|A \cap B|}{|A| + |B|}}$$

$$0 \leq \text{Dice Coefficient} \leq 1$$

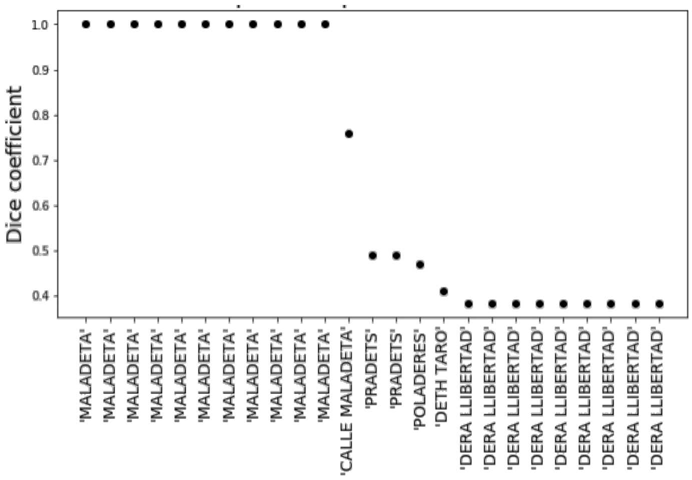


Figure 2: Dice comparison of the bloom filtering results (with regards to the word 'MALADETA')

[2] M. Bellare, et.al; The FFK mode of operation for format preserving encryption, Nist submission, 20, 2010. (https://csrc.nist.gov/csrc/media/projects/block-cipher-techniques/documents/bcm/proposed-modes/ffx/ffx-spec.pdf).

[3] A. Boldyreva, et al; Order-preserving symmetric encryption, Annual International Conference on the Theory and Applications of Cryptografic Techniques, 20, 2009. (https://www.cc.gatech.edu/~aboldyre/papers/bclo.pdf).

[4] Niedermeyer, F. et al; Cryptanalysis of basic bloom filters used for privacy preserving record linkage. Journal of Privacy and Condentiality, 6(2), pp. 59-79. (https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-9-41).

[5] Rair S. et al; *Privacy-preserving record linkage using Bloom filters.* BMC Medical Informatic and Decision Making, 2009 9:41. (http://openaccess.city.ac.uk/14304/).

The same process is also planned to start on M21 for extracting the data of "Enhance control over online banking access" use case.

3.2.5 CRF Datasets progress

The generation of a synthetic dataset for "Production process of aluminium casting" use case has been performed using the IBM TDF tool, according to the definition reported in D2.1. Latest version of data contains 1.2kk entries. Real data for this use case has been generated and it will be updated with the new data.

The preparation of the anonymized datasets related to the "Maintenance and monitoring of production assets" is in progress, as illustrated in Table 1, while the preparation of synthetic data is planned to start on M21.

3.3 Operational experiments definition

The aim of this section is to provide a revised list of the real experiments to be carried out using the I-BiDaaS solution. To this end, first the basic experimental workflow is described. Instances / specialisations of this basic workflow will be considered in the operation plan of each experiment. It should be noted that further to the specific focus of each experiment, as it is expressed through its goals and associated questions reported in sections 3.3.2 - 3.3.4, all experiments will consider the evaluation of usability, operability, robustness, innovation, compliance, privacy, awareness and cost of the I-BiDaaS solution. Although only experiments #3, #4, #5 (batch processing scenario) and #8 will be used for validating the first release of the integrated I-BiDaaS solution (M18), special effort has been made for the initial specification of all experiments.

In addition to experiments which relate to specific industrial sectors, we also discuss the specification of a cross-sectoral experiment, involving multiple data providers, in the subsequent project phases.

3.3.1 Experiment workflow

This section presents the basic workflow for the groups that will participate in the operational experiments using the releases of the I-BiDaaS platform. This workflow has been derived considering the user requirements and system functionality and after discussions with technology providers and end users.

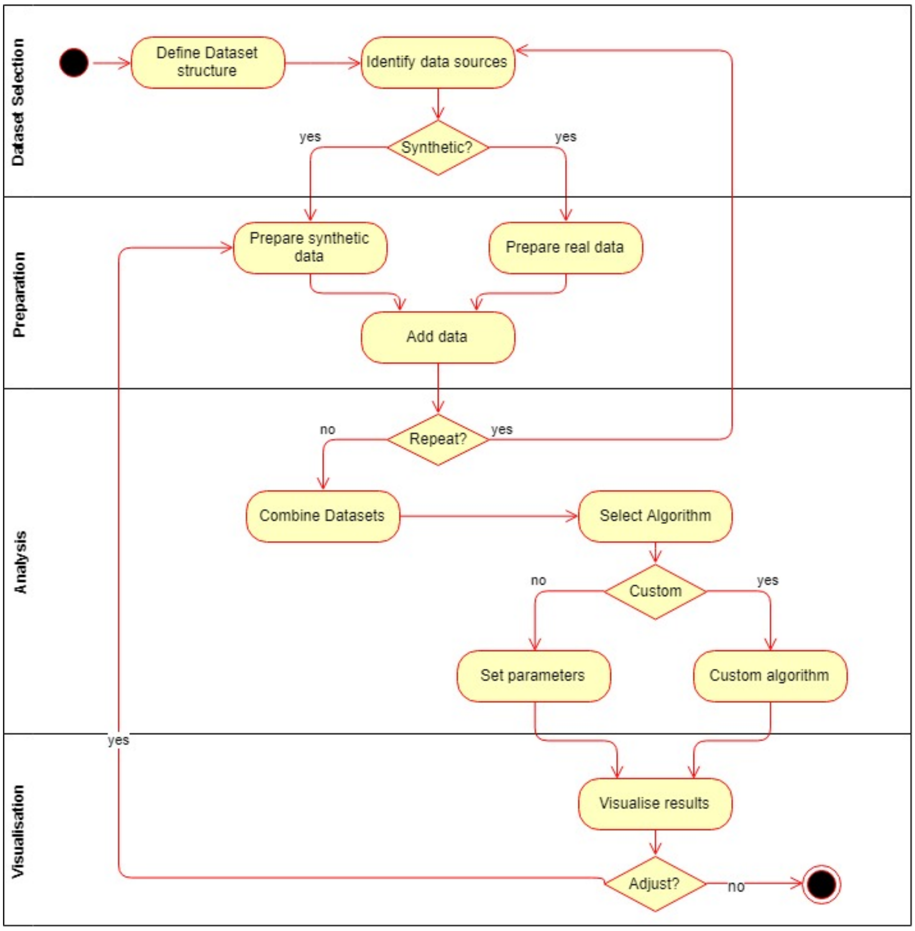


Figure 3: Generic experimental workflow

The core workflow is presented in Figure 3 and describes the steps that users must follow in order to prepare and perform the experiments. In particular, the first step has been the definition of the datasets structure, followed by the identification of multiple data sources to be used for the preparation of the datasets (reported in D2.1). Based on these, the dataset preparation is then performed. This might include the generation of synthetic, realistic data or the aggregation of existing data sources (including both streaming and historical data produced by heterogeneous sources) removing personal identifiable or other sensitive data using appropriate anonymization or tokenization algorithms (as described in sections 3.2.1 and 3.2.2). The aim is to ensure confidentiality of private data while ensuring realistic and meaningful experimentation with high quality test data. Data at this point are added into the batch processing and the streaming analytics modules (as described in deliverables D3.2 and D2.4 respectively). After this, users can combine data and then select the analytics algorithms to perform experiments and extract analytics. Customisation of analytics algorithms is also possible. At the end, visualisation of results helps users review results and explore extracted information to gain meaningful insights and if necessary revise the data fabrication rules and repeat the analysis.

3.3.2 Telecommunication experiments

The following experiment has been specified, during the current reporting period.

<table>
<tr><th>Experiment #3</th><th colspan="3">Employment of bots in call centre <i>(batch processing scenario)</i></th></tr>
<tr><td><b>Experiment's Goals</b></td><td colspan="3">To test I-BiDaaS solution efficiency with respect to the automatic predicting of customer satisfaction.</td></tr>
<tr><td><b>Experiment's Questions</b></td><td colspan="3"><b><i>Q1. What is the quality of the analytics results?</i></b>
Q1.1 How able is the I-BiDaaS platform to detect low customer satisfaction audio calls?
<b><i>Q2. How efficient is the process of data analytics?</i></b>
Q2.1 How many low customer satisfaction audio calls can be detected compared to the ones detected by human agents?
Q2.2 What is the time reduction obtained?
Q2.3 What is the cost reduction obtained?</td></tr>
<tr><td><b>Experimental Workflow <i>(based on the generic workflow, to be further refined)</i></b></td><td colspan="3">1 - Data selection
2 - Real data preparation
- Aggregation of call center data from multiple sources and generation of data files, e.g., audio files, meta-data, etc.
- Execution of ASR model to produce the transcripts
- Execution of speaker segmentation to segment the audio files by the different speakers
- Merging of the ctm (transcripts) and rttm (speakers) files
3 - Data analysis (Apply the NLP model on the input data (merged transcript) to predict the CSI)
4 - Data visualization</td></tr>
<tr><td rowspan="2"><b>Experimental Subjects <i>(participating on any of the steps above)</i></b></td><td><b><i>Role</i></b></td><td><b><i>Steps involved</i></b></td><td><b><i>No of participants</i></b></td></tr>
<tr><td>Data analysts</td><td>1 - 4</td><td>4</td></tr>
</table>

The above experiment has been fully specified and will be used for validating the first release of the integrated I-BiDaaS solution (M18). It should be noted, that two more experiments are under way, corresponding to two more use cases defined during the scoping phase, namely those of "Accurate location prediction with high traffic and visibility" and "Optimization of placement of telecommunication equipment".

3.3.3 Banking experiments

During the current reporting period the following experiments have been defined.

Experiment #4	**Identify non-legitimate connections from loggings of an online banking application** *(batch processing scenario)*		
Experiment's Goals	To test efficiency of I-BiDaaS solution in the context of non-legitimate connections identification from loggings of an online banking application - online banking anomaly detection.		
Experiment's Questions	***Q1. What is the quality of the analytics results?*** Q1.1 How able is the I-BiDaaS platform to detect anomalies? ***Q2. How efficient is the process of data analytics?*** Q2.2 How many potential fraud cases can be solved with I-BiDaaS platform? ***Q3. Does the tokenization/encryption method assure compliance with the current security and privacy regulations?*** ***Q4. Which features can I-BiDaaS provide with regards to other data analytics commercial solutions (such as Data Robot)?***		
Experimental Workflow *(based on the generic workflow, to be further refined)*	1 - Data selection 2- Data preparation – generate rules – fabricate synthetic data – upload data set 3 - Data analysis 4 - Data visualization		
Experimental Subjects *(participating on any of the steps above)*	***Role***	***Steps involved***	***No of participants***
	Quality assurance and control managers	3,4	1
	Data analysts	1,2,3	3
	Infrastructure engineers	1, 2, 3, 4	1
	IT security personnel	1, 3	2

Experiment #5	**Advanced Analysis of bank transfer payment in financial terminal** *(batch processing scenario)*		
Experiment's Goals	To test efficiency of I-BiDaaS solution in the context of anomalies detection in the bank transfers from employees.		
Experiment's Questions	***Q1. What is the quality of the analytics results?*** Q1.1 How able is the I-BiDaaS platform to detect the anomalous bank transfers? ***Q2. How efficient is the process of data analytics?*** Q2.2 How many potential fraud cases can be solved with I-BiDaaS platform? ***Q3. Does the tokenization/encryption method assure compliance with the current security and privacy regulations?*** Q3.1 Does the tokenization/encryption method ensures the privacy of the data for getting out of the premises of CAIXA without business implications? ***Q4. Which features can I-BiDaaS provide with regards to other data analytics commercial solutions (such as Data Robot)?***		
Experimental Workflow *(based on the generic workflow, to be further refined)*	1 - Data selection 2- Data preparation – generate rules – fabricate synthetic data – upload data set 3 - Data analysis 4 - Data visualization		
Experimental Subjects *(participating on any of the steps above)*	***Role***	***Steps involved***	***No of participants***
	Quality assurance and control managers	3,4	1
	Data analysts	1,2,3	3
	Infrastructure engineers	1, 2, 3, 4	1
	IT security personnel	1, 3	2

<table>
<tr><th><b>Experiment #6</b></th><th colspan="3"><b>Analysis of relationships through IP address</b> <i>(batch & streaming processing scenario)</i></th></tr>
<tr><td><b>Experiment’s Goals</b></td><td colspan="3">To validate the use of synthetic data for analysis, if the rules act in the same situations as with the real data.
To test time efficiency of I-BiDaaS solution.</td></tr>
<tr><td><b>Experiment’s Questions</b></td><td colspan="3"><b><i>Q1. Can synthetic data provide the same insights as the real data use case?</i></b>
Q1.1 Has the generated data the same structure as the real one?
Q1.2 How valid is the model generated with synthetic data with regards to the model of the real data?
<b><i>Q2. Is the process of data fabrication more efficient than the process of granting access to real data?</i></b>
Q2.1 How much time (mean) is necessary for generating a volume of synthetic data that can provide a valid model?
Q2.2 How much is the time reduction that we obtain by generating the synthetic data instead of granting permits to an external provider?</td></tr>
<tr><td><b>Experimental Workflow</b> <b><i>(based on the generic workflow, to be further refined)</i></b></td><td colspan="3">1 - Data selection
2- Synthetic data preparation
– generate rules
– fabricate synthetic data
– upload data set
3 – Real data preparation
4 - Data analysis
– select algorithm
– custom algorithm
5 - Data visualization
6 – Adjust data fabrication rules</td></tr>
<tr><td rowspan="5"><b>Experimental Subjects</b> <b><i>(participating on any of the steps above)</i></b></td><td><b><i>Role</i></b></td><td><b><i>Steps involved</i></b></td><td><b><i>No of participants</i></b></td></tr>
<tr><td>Quality assurance and control managers</td><td>3,4</td><td>1</td></tr>
<tr><td>Data analysts</td><td>1,2,3</td><td>3</td></tr>
<tr><td>Infrastructure engineers</td><td>1, 2, 3, 4</td><td>1</td></tr>
<tr><td>IT security personnel</td><td>1, 3</td><td>2</td></tr>
</table>

3.3.4 Manufacturing experiments

This section reports on the progress of the CRF experiments definition (use case by use case).

Experiment #7	**Maintenance and monitoring of production assets** *(batch processing scenario)*		
Experiment's Goals	To test efficiency of I-BiDaaS solution in the context of anticipation of maintenance events (alarm).		
Experiment's Questions	***Q1. What is the quality of the analytics results?*** Q1.1 What is the accuracy of new models with respect to internal CRF models in use (geographical representation of the process)? ***Q2. How efficient is the process of data analytics?*** Q2.2 How efficient is the performance of the analytics application (algorithm)?		
Experimental Workflow *(based on the generic workflow, to be further refined)*	1 - Data selection 2- Synthetic data preparation – generate rules – fabricate synthetic data – upload data set 3 – Real data preparation 4 - Data analysis – select algorithm – custom algorithm 5 - Data visualization 6 – Adjust data fabrication rules		
Experimental Subjects *(participating on any of the steps above)*	***Role***	***Steps involved***	***No of participants***
	Quality assurance and control managers	1, 4	1
	Data analysts	1- 4	2
	Financial administrators	4	1
	Infrastructure engineers	1	1
	IT security personnel	1	1

<table>
<tr><th>Experiment #8</th><th colspan="3">Production process of aluminium casting <em>(batch processing scenario)</em></th></tr>
<tr><td><strong>Experiment’s Goals</strong></td><td colspan="3">To test the efficiency of I-BiDaaS solution in the context of correlating defects with the production process parameters.</td></tr>
<tr><td><strong>Experiment’s Questions</strong></td><td colspan="3"><strong><em>Q1. What is the quality of the analytics results?</em></strong>
Q1.1 What is the accuracy of new models with respect to internal CRF Aluminium Casting models?
<strong><em>Q2. How efficient is the process of data analytics?</em></strong>
Q2.2 How efficient is the performance of the analytics application (algorithm)?</td></tr>
<tr><td><strong>Experimental Workflow <em>(based on the generic workflow, to be further refined)</em></strong></td><td colspan="3">1 - Data selection
2- Synthetic data preparation
– generate rules
– fabricate synthetic data
– upload data set
3 – Real data preparation
4 - Data analysis
– select algorithm
– custom algorithm
5 - Data visualization
6 – Adjust data fabrication rules</td></tr>
<tr><td rowspan="6"><strong>Experimental Subjects <em>(participating on any of the steps above)</em></strong></td><td><strong><em>Role</em></strong></td><td><strong><em>Steps involved</em></strong></td><td><strong><em>No of participants</em></strong></td></tr>
<tr><td>Quality assurance and control managers</td><td>1, 4</td><td>1</td></tr>
<tr><td>Data analysts</td><td>1-4</td><td>2</td></tr>
<tr><td>Financial administrators</td><td>4</td><td>1</td></tr>
<tr><td>Infrastructure engineers</td><td>1</td><td>1</td></tr>
<tr><td>IT security personnel</td><td>1</td><td>1</td></tr>
</table>

3.3.5 Cross-sectoral experiment definition

The combination of different data-sources is a challenging task for many reasons. First, the amount of data can become a bottleneck when need to transfer data from different locations. Second, in many cases the data may have different policy and access controls that prohibit the outsourcing. Third, data may spread to different data silos (e.g., different database vendors, setups, etc.), and have different formats.

A cross-sectorial experiment requires a general solution that can address the challenges mentioned above and is able to support the cross-sectorial use cases that may be defined. For this reason, we define two possible scenarios, regarding where the data are stored:

1) *Data is copied over to a common database.* This is the classic data-warehouse solution, and ensures a common, usually pre-defined, format. The disadvantages though, are the data security and the possibility that data become stale and out-of-sync. In particular, the company’s data now resides in a database that is hosted outside its premises; as such it is necessary to periodically update and synchronize. However, in several scenarios, such as in our financial pilot, this is prohibited by regulations, security policies, etc. As such, the

experimental protocol for this case should include either the proper anonymization of any sensitive or secret data either the fabrication of the data before leaving the premises.

2) *Data is combined and processed in run-time.* This is the classic stream processing approach, in which data are generated at real-time and due to their volume and/or velocity can only be processed in-memory. The disadvantages of this approach are that it cannot support complicated queries (e.g., joins) and can process data only for specific time windows. Again, the data should be properly anonymized or fabricated in case it is leaving the company's premises. As such, the experimental protocol for this case should also include these two tasks.

Finally, the different data sources should be presented via the data virtualization environment, enabling a data analyst to model relationships between data sets as though they were physically present in the same data silo. The interface will also generate a query execution path that identifies the optimal way to combine different data sources, either by copying the data over a common database, either by combining and processing the data sets at run-time.

4 Experimental Evaluation

4.1 Overview

The aim of this section is to provide a brief overview of the I-BiDaaS first complete prototype and report on the progress of the evaluation process, considering the evaluation methodology defined in section 2.

4.2 The first release of the integrated I-BiDaaS solution

Following the delivery of the I-BiDaaS Minimum Viable Product (MVP) in M12 (see Del. 5.2), this section provides an updated overview of the I-BiDaaS integrated solution and its components.

Following the recent developments, the I-BiDaaS solution will be based on two project types (batch and streaming) and is supported by a simple authentication process. The "Analysis of relationships through IP address" from CAIXA (the MVP use case), is used to illustrate the data flow in both project types.

In particular, in the batch processing subcase, the data flow is the following:

- A file of synthetic data in SQLite format is created by TDF. It is a single table database with the structure presented in I-BiDaaS Deliverable 2.1.
- The file is imported in a central Cassandra Database as a Cassandra table with the same structure.
- Hecuba runs on the table and calculates the user relations between people. Results (pair of user ids) are stored in a new table in Cassandra and a file in CSV format.

In the stream processing subcase, the data flow is the following:

- Random transactions of users (from the pool of user IDs created in the first synthetic dataset) are created, in the format depicted in I-BiDaaS Deliverable 2.1. Each transaction is published as a simple string with comma separated values via MQTT in a topic called "caixa_transaction_pair" in UM.
- An APAMA[6] application:
    - loads the user relations (groups) that were created in the batch processing subcase in memory
    - subscribes to "caixa_transaction_pair" topic and checks every incoming transaction to see if the pair of users are in a group sends every transaction that is between a "related" pair of users as a new message to another topic called "caixa_transaction_related".

The overall architecture of the I-BiDaaS solution has not been modified since the one presented for the MVP, in Deliverable 5.2. The revised dataflow is presented in the following diagram (Figure 4).

[6] *Apama Streaming Analytics* is a Complex Event Processing (CEP) and Event Stream Processing (ESP) engine, developed by Software AG. Apama serves as a platform for performing streaming analytics over a range of high volume/low latency inputs and applications, such as IoT devices, financial exchanges, fraud detection, social media and similar

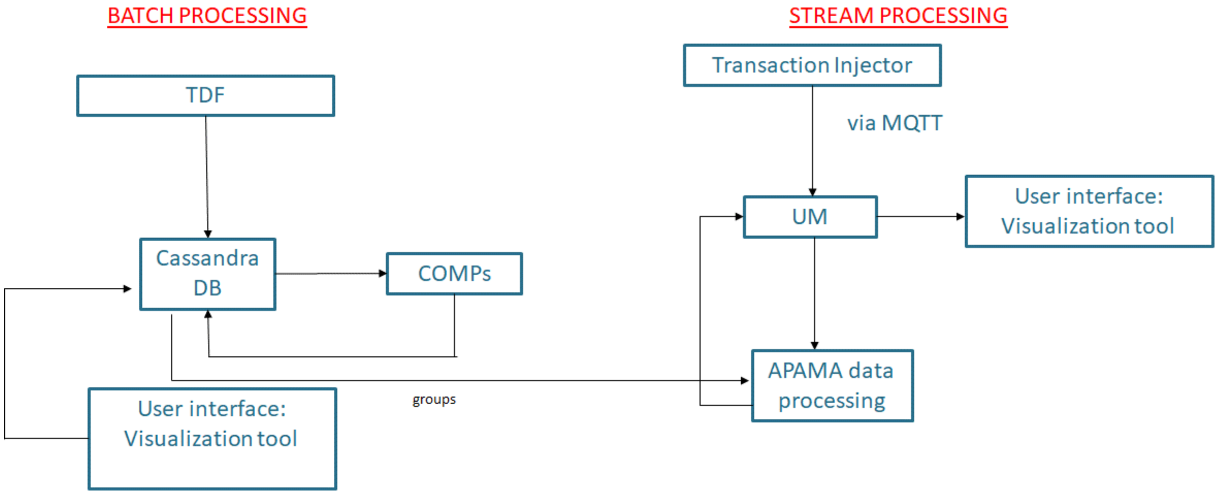


Figure 4: Revised data flow

Regarding the updates on the provision and configuration of Infrastructure Resources, the I-BiDaaS integrated solution is based on 4 working environments:

- MVP M12 managed directly on the virtual environment
- Private all-in-one CSP (OpenStack based)
- Private multi-node CSP
- Commodity Cluster

With respect to the updated datasets and use cases to be part of the I-BiDaaS integrated solution, the following Table 4 summarizes the datasets, the data types and the use case owners respectively:

Table 4: Datasets of the I-BiDaaS integrated solution

No.	I-BiDaaS dataset	Type of Data	Partner owner
3	Call Center Data	Real data (In-house)	TID
4	Bank Transfer	Tokenized	CAIXA
5	IP Address	Synthetic	CAIXA
9	Aluminium Casting	Synthetic & Anonymised	CRF

The hardware specifications for the aforementioned use cases and data sets are summarized in the following Table 5.

Table 5: Hardware specification

Component(s)	Platform	Cores	RAM	Storage
IBM InfoSphere® Optim Test Data Fabrication	Linux server (x86_64)	4	8GB	20GB
Hecuba (2 nodes)	Linux server (x86_64)	4 per node	4GB per node	20GB per node
Software AG APAMA Server	Linux server (x86_64)	4	8GB	50GB
Software AG Universal Messaging	Linux server (x86_64)	4	8GB	50GB
Visualisation Toolkit / I-BiDaaS Dashboard	Linux server (x86_64)	2	4GB	30GB
Database (Cassandra)	Linux server (x86_64)	4	4GB	100GB
Integration Tools	Linux server (x86_64)	4	16GB	100GB
	TOTAL	30	56 GB	400 GB

Finally, regarding visualisation, the following procedures have been defined for the I-BiDaaS integrated solution:

Data selection

- Fabrication: 9 TDF project files (one per use case) accepting predefined parameters
- Orchestrator API is used to store user's selected parameters in the internal I-BiDaaS DB (MySQL)
- Fabrication starts from the interface by calling the relevant TDF REST service
- Dataset Selection: user selects the dataset by choosing a datafile existing in I-BiDaaS filesystem or by pointing at a keyspace in Cassandra
- A preview of the selected data is displayed

Analysis

- Upon algorithm selection the relevant parameters are supplied by the orchestration API and filled-in in the interface. Experiment starts and execution statistics are retrieved by log files supplied by the cloud management software.
- Multiple experiments (different algorithms/params on same dataset) can be run at any given time and are visible when viewing the project
- Custom code can be uploaded as a zip file and executed. Zip templates and instructions will be available

Visualisation

- One Predefined visualisation available per use case
- 'Create own visualisation' link that opens up Mashzone passing a pointer (e.g. table name) to the experiment results.

The following figures provide the current status with respect to the aforementioned flows for the batch processing and the stream processing cases respectively.

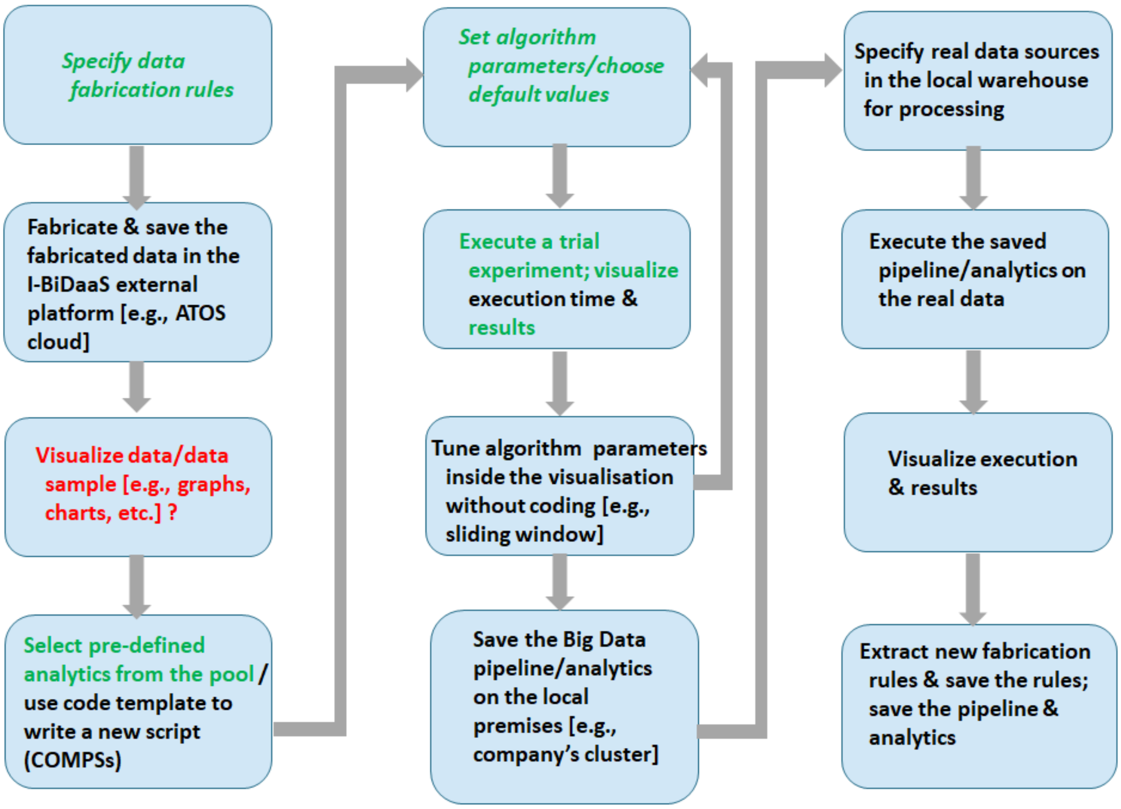

Specify data fabrication rules
Fabricate & save the fabricated data in the I-BiDaaS external platform [e.g., ATOS cloud]
Visualize data/data sample [e.g., graphs, charts, etc.] ?
Select pre-defined analytics from the pool / use code template to write a new script (COMPSs)
Set algorithm parameters/choose default values
Execute a trial experiment; visualize execution time & results
Tune algorithm parameters inside the visualisation without coding [e.g., sliding window]
Save the Big Data pipeline/analytics on the local premises [e.g., company's cluster]
Specify real data sources in the local warehouse for processing
Execute the saved pipeline/analytics on the real data
Visualize execution & results
Extract new fabrication rules & save the rules; save the pipeline & analytics

Figure 5: The batch processing flow

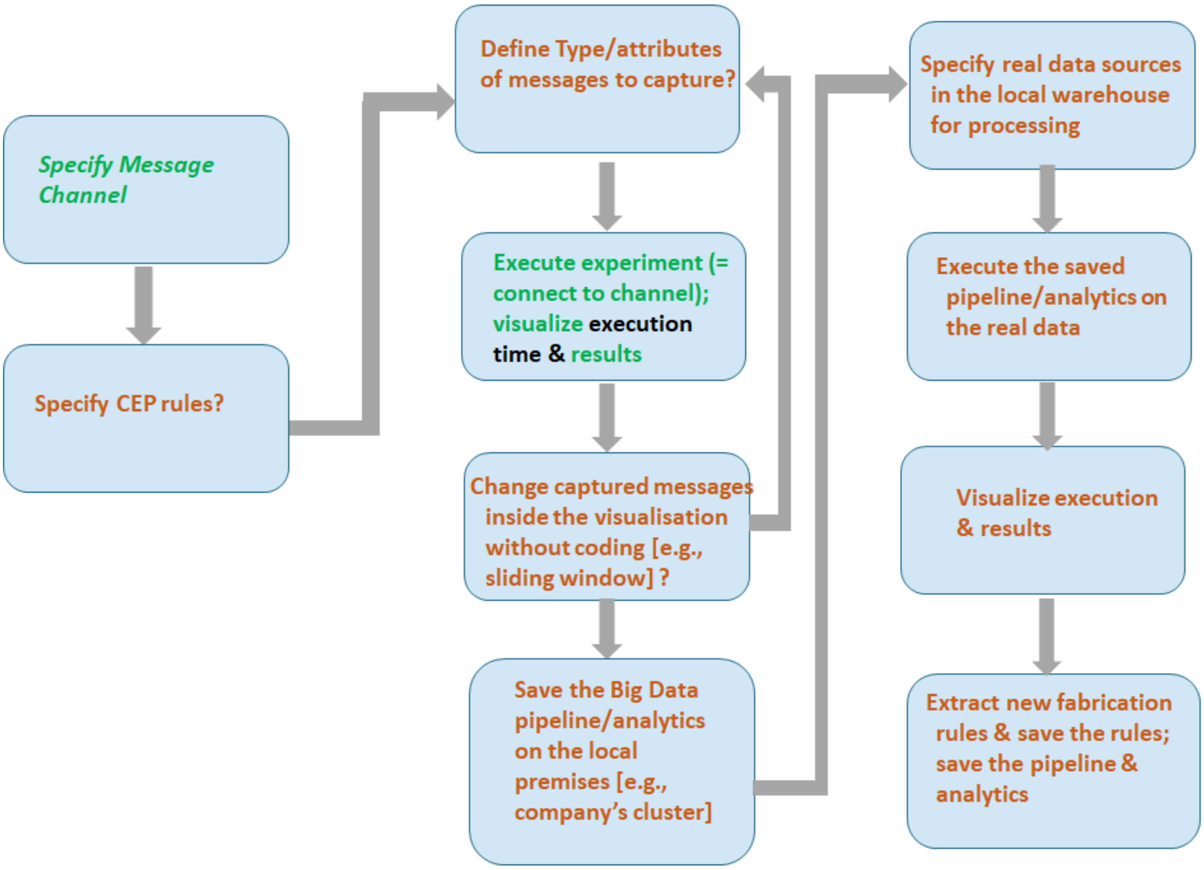
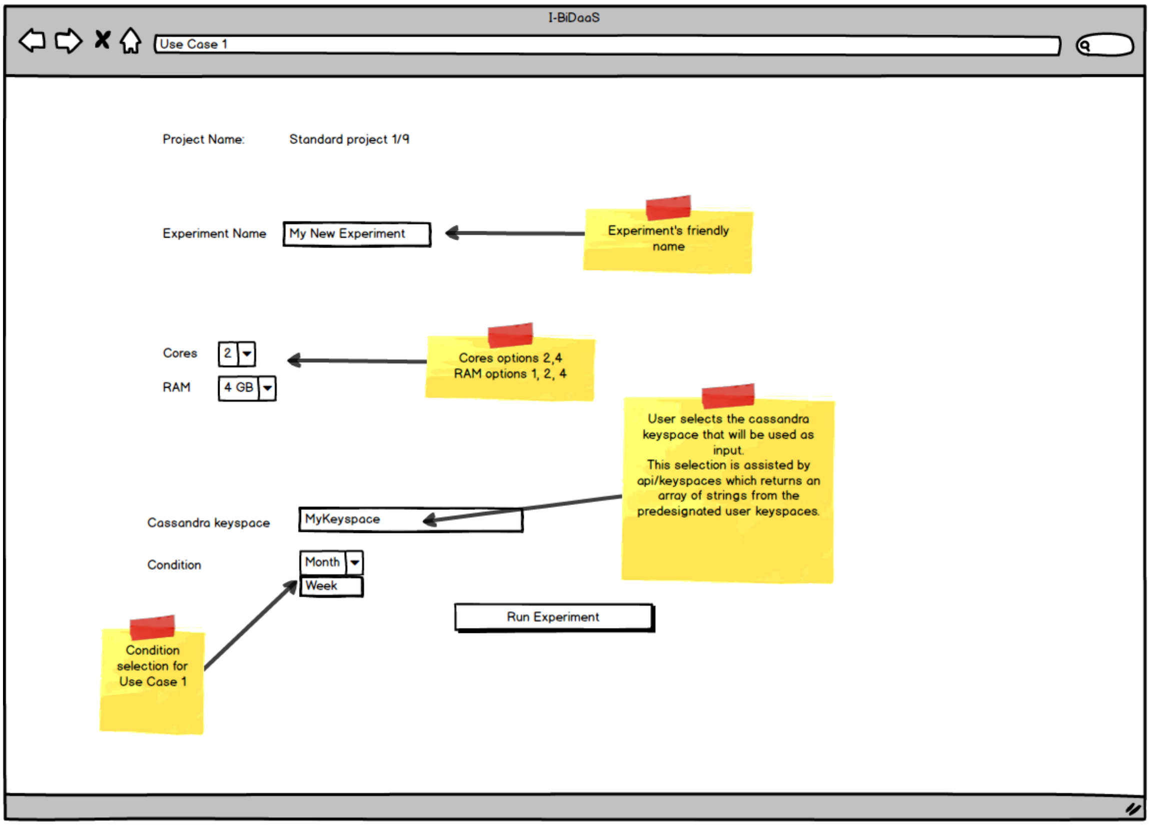

Specify Message Channel
Specify CEP rules?
Define Type/attributes of messages to capture?
Execute experiment (= connect to channel); visualize execution time & results
Change captured messages inside the visualisation without coding [e.g., sliding window] ?
Save the Big Data pipeline/analytics on the local premises [e.g., company's cluster]
Specify real data sources in the local warehouse for processing
Execute the saved pipeline/analytics on the real data
Visualize execution & results
Extract new fabrication rules & save the rules; save the pipeline & analytics

Figure 6: The streaming processing flow

4.2.1 Wireframes of the I-BiDaaS solution for the functionality of standard/default experiment

In order to demonstrate and validate the I-BiDaaS solution, a set of two complementary generic use cases have also been deployed. The following wireframes display the functionality for standard/default experiment, custom project creation and custom experiment and experiment for use cases 1 and 2.



Project Name: Standard project 1/9

Experiment Name: My New Experiment

Experiment's friendly name

Cores: 2

RAM: 4 GB

Cores options 2,4
RAM options 1, 2, 4

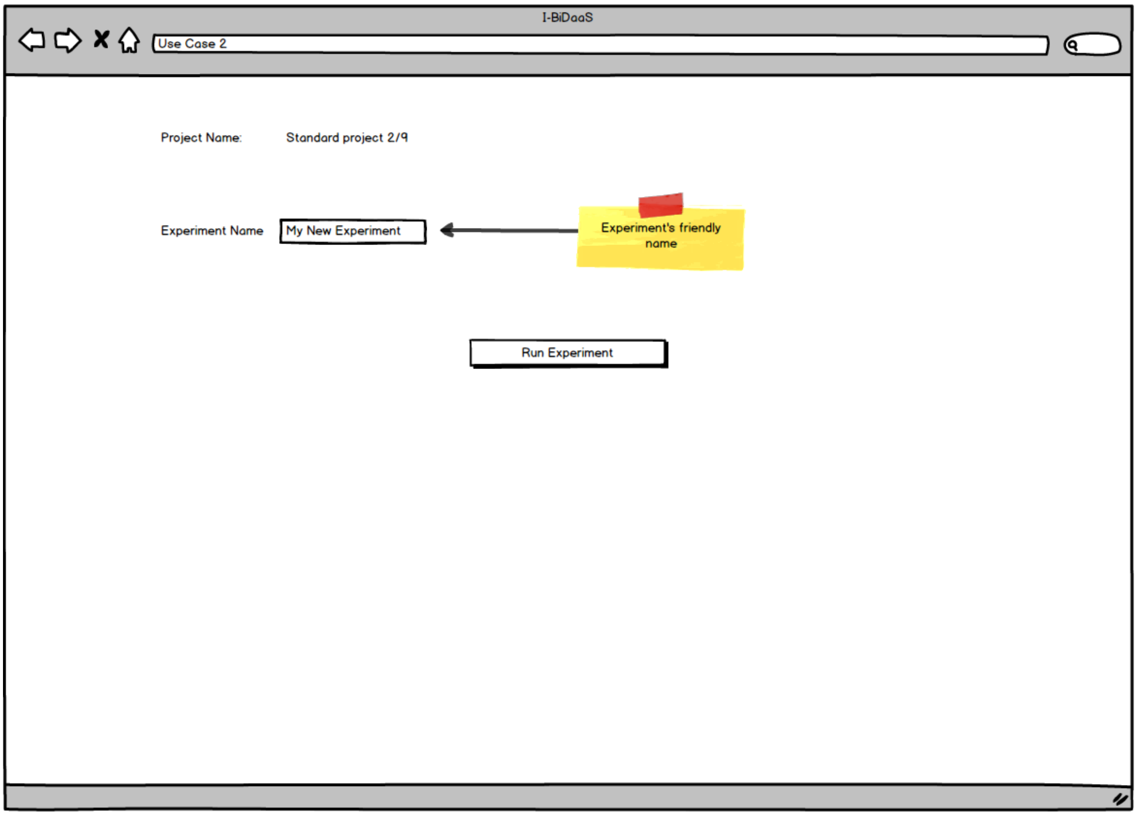
User selects the cassandra keyspace that will be used as input.
This selection is assisted by api/keyspaces which returns an array of strings from the predesignated user keyspaces.

Cassandra keyspace: MyKeyspace

Condition: Month
Week

Condition selection for Use Case 1

Figure 7: Standard/default experiment for use case 1

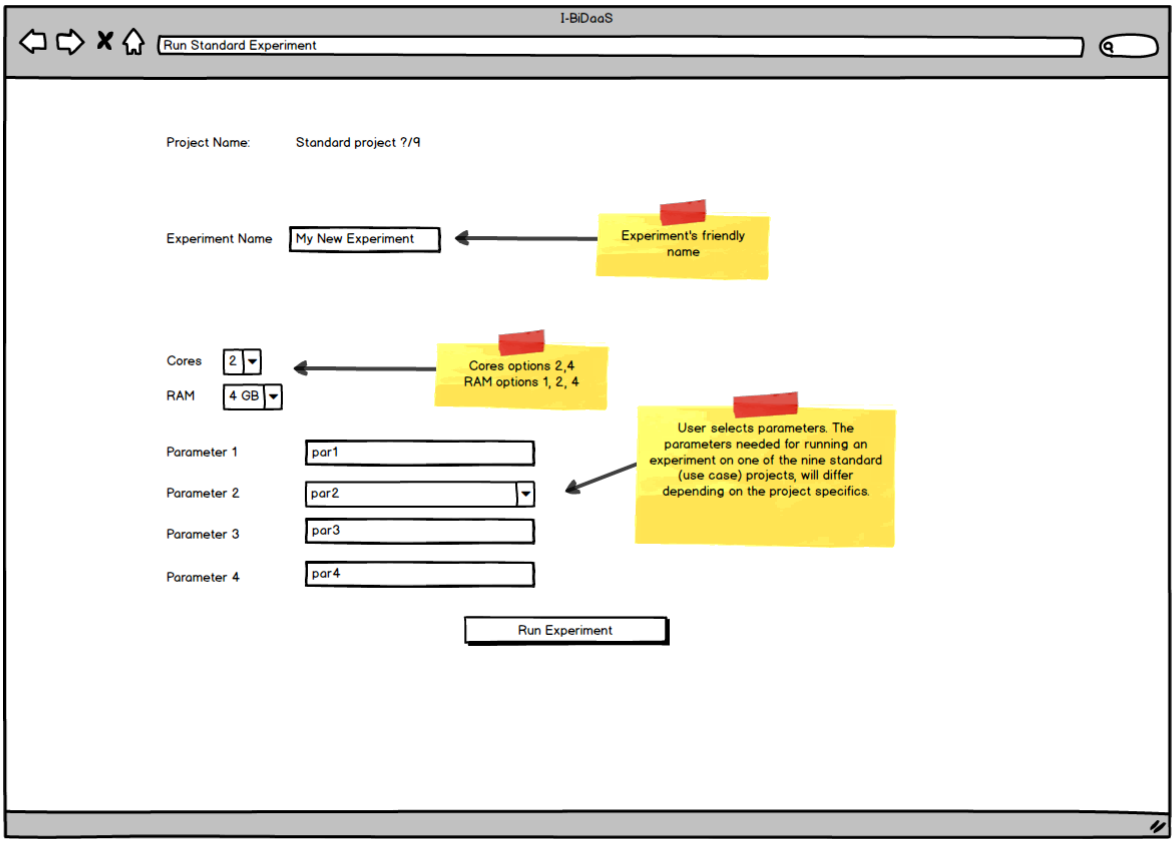


Project Name: Standard project 2/9

Experiment Name: My New Experiment

Experiment's friendly name

Figure 8: Standard/default experiment for use case 2



Project Name: Standard project ?/9

Experiment Name: My New Experiment

Experiment's friendly name

Cores: 2

RAM: 4 GB

Cores options 2,4
RAM options 1, 2, 4

Parameter 1: par1

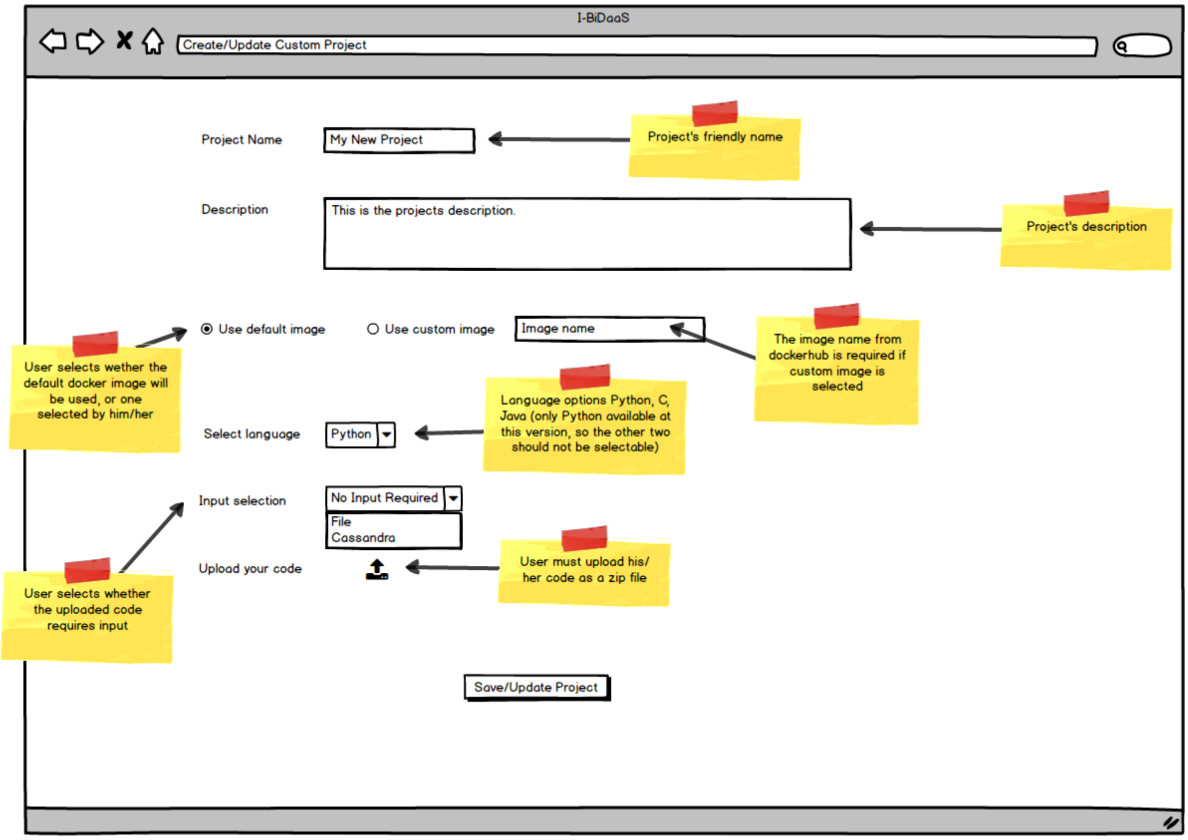
Parameter 2: par2

Parameter 3: par3

Parameter 4: par4

User selects parameters. The parameters needed for running an experiment on one of the nine standard (use case) projects, will differ depending on the project specifics.

Figure 9: Run standard experiment



Project Name: My New Project

Project's friendly name

Description: This is the projects description.

Project's description

● Use default image ○ Use custom image Image name

User selects wether the default docker image will be used, or one selected by him/her

The image name from dockerhub is required if custom image is selected

Select language: Python

Language options Python, C, Java (only Python available at this version, so the other two should not be selectable)

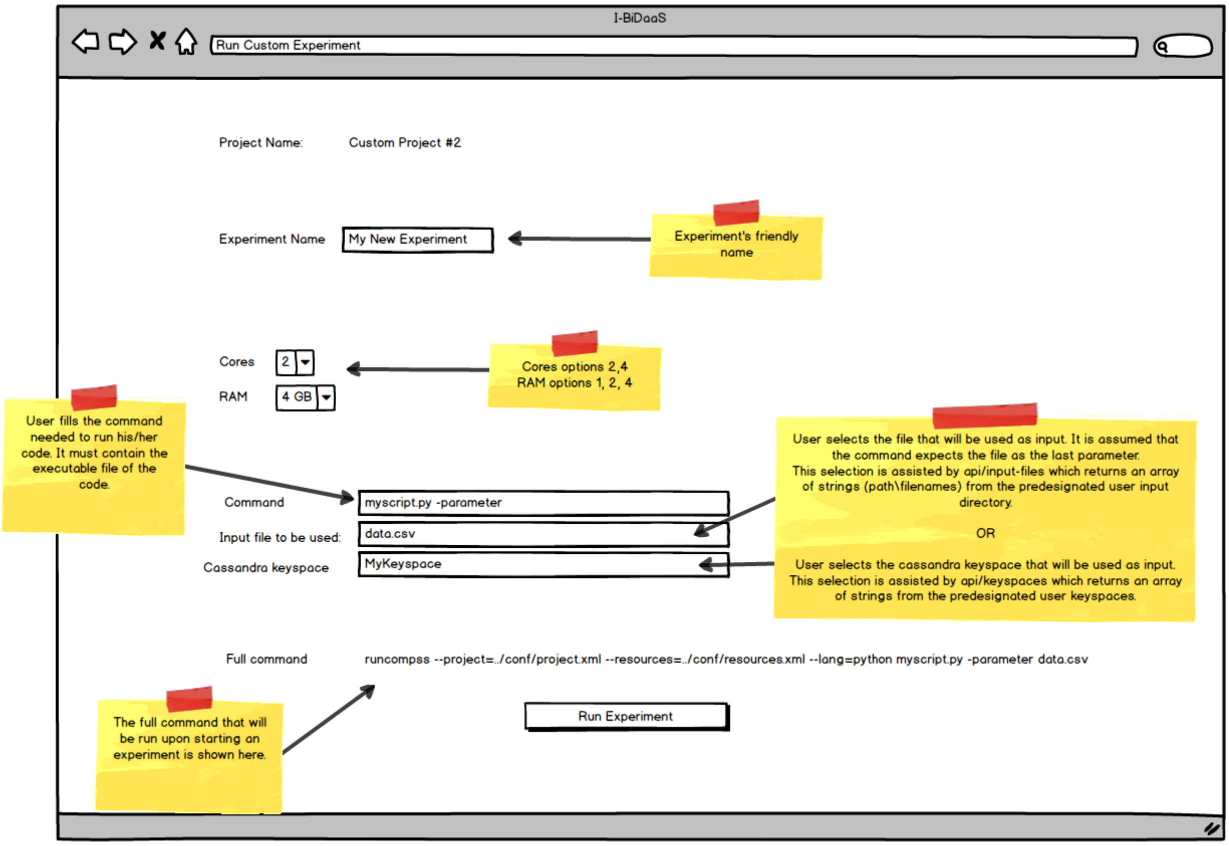
Input selection: No Input Required
File
Cassandra

User selects whether the uploaded code requires input

Upload your code

User must upload his/her code as a zip file

Figure 10: Create/Update custom project



Project Name: Custom Project #2

Experiment Name: My New Experiment

Experiment's friendly name

Cores: 2

RAM: 4 GB

Cores options 2,4
RAM options 1, 2, 4

User fills the command needed to run his/her code. It must contain the executable file of the code.

Command: myscript.py -parameter

Input file to be used: data.csv

Cassandra keyspace: MyKeyspace

User selects the file that will be used as input. It is assumed that the command expects the file as the last parameter.
This selection is assisted by api/input-files which returns an array of strings (path\filenames) from the predesignated user input directory.

OR

User selects the cassandra keyspace that will be used as input.
This selection is assisted by api/keyspaces which returns an array of strings from the predesignated user keyspaces.

Full command:

```
runcompss --project=../conf/project.xml --resources=../conf/resources.xml --lang=python myscript.py -parameter data.csv
```

The full command that will be run upon starting an experiment is shown here.

Figure 11: Run custom experiment

4.3 Experiments verification and validation

The aim of this section is to provide an overview of the evaluation indicators as well as the industrial benchmarks that will be used for the verification and validation of the I-BiDaaS solution in parts and as a whole (quantitative evaluation). The discussion focuses on the current implementation of the solution, in the form of the first complete prototype. In addition, we provide an update of indicators to be used for the experimental validation (qualitative evaluation), in the real work setting with the participation of industrial users (as described in sections 3.3.2 - 3.3.4).

4.4 Quantitative evaluation

As already mentioned, the quantitative (technical) evaluation focuses on the evaluation of quality of the I-BiDaaS platform in parts and as a whole through testing using appropriate benchmarks, where available.

The following table is an updated list of indicators to be used for the testing of each platform component:

Table 6: Component evaluation

Indicator	Metric	Benchmark (if applicable)[7]
TDF (IBM)		
Scalability	Linear (in number of TDF instances) speedup in generated records	
Validity	Generated data must fit the data model	
Performance	Number of generated records per time unit	
Accuracy	Measured against real data	Benchmark the quality of the fabricated data by applying analytics on both real and fabricated data.
Availability	No of crashes	
Universal Messaging (SAG)		
Scalability	Response time	seconds
	Data throughput	MB/second
	Resource utilisation	MB and CPU %
Operational performance	Execution time	seconds
	Latency	seconds
Reliability	Data failure	
Compliance	Measured against relevant standards	JMS, MQTT, AMQP
Advanced ML (UNSPMF)		
Quality of results	For optimization solver: Objective function value; constraints violation value	Standard solver on the same problem, if feasible for the given data set and if implementation is available: CVXPY[8], relevant MPI implementation. Targeted to have a small deviation with respect to the benchmark, e.g., 10^(-2)
	For supervised learning: Training error; testing error; classification accuracy	Benchmark: Sklearn[9], for moderate size data. Target: have a comparable result with the benchmark, with improved scalability.
	For clustering: silhouette score	Benchmark: Sklearn, for moderate size data. Target: have a comparable result with the benchmark, with improved scalability.
Scalability	Execution time versus number of nodes (cores	Respective MPI implementation, if available. Target: have scalability which is comparable with MPI (lower performance than MPI expected due to using COMPSs framework with less programming and system optimization effort.)

[7] For the non-applicable: Units are included instead of benchmarks.

[8] Python-embedded modelling language for Convex optimization problems, available at https://www.cvxpy.org/

[9] https://scikit-learn.org/stable/index.html

Performance	For testing a novel algorithm: Number of iterations; number of messages exchanged between the nodes	Relevant state of the art algorithm in the literature.
CEP Engine (SAG)		
Scalability	Response time	seconds
	Data throughput	MB/seconds
	Resources Utilisation	MB and CPU %
GPU-accelerated pattern matching (FORTH)		
Performance	Throughput	Number of tuples processed per second, and speed-up achieved over a CPU-only implementation.
	Latency	Time required to process a specific number of tuples and speed-up achieved over a CPU-only implementation.
Hecuba DBS (BSC)		
Scalability	Speedup	Benchmark: Dislib[10] Baseline: Dislib using files
Operational Performance	Response time	
	IOPS	
	Disk usage	
Availability	% timeouts	
Reliability	Fault tolerance on down nodes (amount of missed data and slowdown during the recovery process)	
Qbeast (BSC)		
Effect on Machine learning algorithms	Speedup	Benchmark: Dislib and MLlib[11] Baseline: Dislib and MLlib without using Qbeast
Scalability		Benchmark: synthetic geographical queries Baseline: PostGis[12]
Operational performance	Response time	
	IOPS	
	Disk usage	
Availability	% timeouts	
Reliability	Fault tolerance on down nodes (amount of missed data and slowdown during the recovery process)	
Resource management and orchestration module (ATOS)		
Operational performance	Response time	- Page response time (in seconds) - Transactions processed - Network error, latency & utilization

[10] The Distributed Computing Library, available at https://dislib.bsc.es/

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

[12] A spatial database extender for PostgreSQL object-relational database, available at https://postgis.net/

	Resource utilization	Delay between a client request and a cloud service provider's response. (in seconds)
Scalability	Average of assigned resources among the requested resources	Makespan of the service creation until the deployment of the resources are acknowledged. (in seconds)
Availability	Responsiveness	Verify that the number of resources is between the resource limits (max & min) as defined in the blueprint.
Reliability	Service Constancy	How much time the service provider guarantees that your data and services are available (in percentage)
	Accuracy of Service	Rules defined to ensure service reliability (number of replicas and policy types)
	Fault Tolerance	Ability to continue providing service after a failure
Visualisation Tool (AEGIS)		
Operational performance	Response time	7-10s
Availability	Uptime	95%
Reliability	Fault Tolerance	Interface responsive in case of data errors. Informative messages to users.
Usability (Efficiency)	Task time efficiency	>20% Decrease with respect to current times
	Perception of time required to accomplish a task	>30% Decrease with respect to current times
	Perception of task completion quality	>80% positive perception. Results via 1-5 scaled questions.
Usability (Satisfaction)	Degree to which user needs are satisfied - look and feel	>80% positive perception. Results via 1-5 scaled questions.

The following table is an updated list of indicators to be used for the testing of the integrated solution:

Table 7: Integrated solution evaluation

Indicator	Metric	Benchmark
Usability	Task time efficiency	
	Perception of time required to accomplish a task	
	Perception of task completion quality	
Scalability	Speedup	
Operational Performance	Response time (Latency)	
	Data throughput (IOPS, no of generated data records per time unit)	
	Resources utilization (storage, memory, CPU)	
Availability	Uptime, % timeout	
Reliability	Data failure, Fault tolerance	

Data Security	Compliance with relevant security and privacy regulations and standards	
Privacy	Compliance with relevant security and privacy regulations and standards	
Compliance	Measured against relevant standards	
Cost	Compared against commercial alternatives	

The collection of quantitative data regarding the quality and especially the performance of the I-BiDaaS integrated solution, will also involve the use of applicable benchmarking tools (see Table 8). Different benchmarks are applicable for different data types and metrics measured. More information regarding the proposed benchmarking process is reported in the experimental protocol (D1.3, section 5.6), as well as in D5.2 (section 2.6).

Table 8: Classification of applicable Big Data benchmarking tools[13]

Category	Year	Name	Type	Domain	Data Type	Metric
Micro Benchmarks	2010	HiBench[14]	Micro-benchmark Suite	Micro-benchmarks, Machine Learning, SQL, Websearch, Graph, Streaming Benchmarks	Structured, Text, Web Graph	Execution time, throughput
	2015	SparkBench[15]	Micro-benchmark Suite	Machine Learning, Graph Computation, SQL, Streaming Application	Structured, Text, Web Graph	Execution time, throughput
	2013	BigDataBench[16]	Micro-benchmark Suite	Online service, offline analytics, graph analytics, artificial intelligence (AI), data warehouse, NoSQL and streaming	Graph, Network, Text, NLP, Web, Image, Audio, Spatio-Temp. Time Series, IoT, Structured, BI	Performance

[13] Adjusted from "**DataBench 2018**, Deliverable D1.1 Industry Requirements with benchmark metrics and KPIs, Report, available at https://www.databench.eu/wp-content/uploads/2019/01/databench-d1.1-ver.1.0.pdf"

[14] HiBench Suite, https://github.com/intel-hadoop/HiBench

[15] SparkBench, https://github.com/CODAIT/spark-bench

[16] BigDataBench. http://prof.ict.ac.cn/BigDataBench.

	2010	YCSB[17]	Micro-benchmark	Cloud OLTP operations	Structured	Execution time, throughput
	2017	TCPx-IoT[18]	Micro-benchmark	Workloads on typical IoT Gateway systems	Structured, IoT	Performance
Application Benchmarks	2015	Yahoo Streaming Benchmark[19]	Application Streaming Benchmark	Advertisement analytics pipeline	Structured, Time Series	Execution time, throughput
	2013	BigBench/TPCx-BB[20]	Application End-to-end Benchmark	A fictional product retailer platform	Structured, Text, JSON logs	Performance
	2017	BigBench V2[21]	Application End-to-end Benchmark	A fictional product retailer platform	Structured, Text, JSON logs	Performance
	2018	ABench[22] (Work in Progress)	Big Data Architecture Stack Benchmark	Set of different workloads	Structured, Text, JSON logs	

The initial results of the quantitative evaluation for the MVP are reported in D5.2. Additional data regarding the quantitative evaluation of the current implementation (M18) of platform components are reported in D2.4 and D3.2.

[17] YCSB, https://github.com/brianfrankcooper/YCSB

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

[19] YSB, https://github.com/yahoo/streaming-benchmarks

[20] BigBench, https://github.com/intel-hadoop/Big-Data-Benchmark-for-Big-Bench

[21] Ahmad Ghazal et al. 2017: BigBench V2: The New and Improved BigBench. ICDE 2017: 1225-1236

[22] Todor Ivanov 2018, Rekha Singhal: ABench: Big Data Architecture Stack Benchmark. Companion of the 2018 ACM/SPEC International Conference on Performance Engineering, ICPE 2018, Berlin, Germany, April 09-13, 2018

4.5 Qualitative evaluation

As discussed in section 3, the definition of the experimental qualitative indicators to be measured, are guided by the experiments questions and correspond to the actual data that will be collected during each experiment. In particular, we differentiate between: business KPIs that measure the impact of technology on the business domain and indicators that measure the performance of the I-BiDaaS solution at application and platform level. In this way, we make more explicit the alignment between business impact and perceived platform quality in the context of each experiment. The following sections 4.5.1 - 4.5.3 describe the indicators that are expected to be measured during the experiments specified in section 3.3, to be further refined during the experimentation phase of I-BiDaaS.

4.5.1 Telecommunication experiments

Experiment #3 - Employment of bots in call centre

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

4.5.2 Banking experiments

Experiment #4 - Identify non-legitimate connections from loggings of an online banking application

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

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

	Indicator	Metric	Baseline value / Benchmark	expected I-BiDaaS value
Business KPIs	Time efficiency Cost reduction Data accessibility	Time to access data Infrastructure cost Number of people accessing data	2 weeks - 1 month Internal temporal storage cost Order of magnitude of 10	
Application Level Performance indicators	End-to-end execution time	Data charging time Time to get analytics results. Time to generate business rules.	minutes minutes 1 week	
	Accuracy and reliability of the analytical process	Anomalies detected	Domain specific evaluation / Number of Anomalies extracted with commercial product (Data Robot)	
		Confusion matrix, TP, TN, etc.	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.	
Platform level performance indicators	Cost	Price of technologies.	Cost of commercial product licenses (e.g. DataRobot).	

Experiment #6 - Analysis of relationships through IP address

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

4.5.3 Manufacturing experiments

Experiment #7 - Maintenance and monitoring of production assets

	Indicator	Metric	Baseline value / Benchmark	expected I-BiDaaS value
Business KPIs	Product / Service quality Cost reduction	OEE JPH Maintenance cost	94% 18 200 kEUR every 3 months	94.5 % 18.2 100 kEUR every 3 months
Application Level Performance indicators	Execution Time Data Quality	Time to produce decisions Accuracy of new models with respect to internal CRF models in use	1 month ND	Near real time 20%
Platform level performance indicators	Cost	Cost regarding personnel time spent (for analysis process) E.g. time spent for data anonymization	30 kEUR	TBD (platform cost)

Experiment #8 - Production process of aluminium casting

	Indicator	Metric	Baseline value / Benchmark	expected I-BiDaaS value
Business KPIs	Product quality	Quality control 1 Quality control 2 Quality control 3	67% 27% 6%	77 % 22 % 1 %
Application Level Performance indicators	Execution Time	Time to produce automated decisions	1 month	Near real time
	Data Quality	Accuracy of new models with respect to internal CRF Aluminium Casting models	ND	20%
Platform level performance indicators	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization	30 kEUR	TBD (platform costs)

4.5.4 Common experimental indicators

Regardless the specific goals of each experiment, all experiments will consider the evaluation of usability, operability, robustness, innovation, compliance and privacy awareness and cost of the I-BiDaaS solution. To this end participants will be asked to answer relevant questions. An indicative list of such questions is presented below:

Participants question list
Usability
How easy has it been was to perform each task (Data selection, preparation, analysis, visualisation)?
Operability
Can the solution be put into practice the real business setting?
Robustness
Will the experiment bring the same results if the experimental conditions change?
Innovation
Which features can I-BiDaaS provide with regards to other data analytics commercial solutions in use?
Compliance
Does the I-BiDaaS solution assure compliance with the current security and privacy regulations?
Privacy Awareness
Does the I-BiDaaS solution ensure the privacy of the data for getting out of company premises without business implications?

Cost
To what extent is the I-BiDaaS solution cost effective? Consider the following dimensions: ***Licence cost / support (personnel) cost / infrastructure cost / encryption cost, compared to current solution in use***

5 Impact Analysis

This section discusses the results of the Innovation Phase (M8-M18) of the project and provides an analysis with respect to the expected project level innovation and achievements. In addition, it describes the implemented or prospective activities (*EAB meetings / I-BiDaaS workshops, hackathons) aiming to demonstrate the I-BiDaaS solution and involve external users into the evaluation process.*

5.1 Progress report

From the very beginning of the I-BiDaaS project, the main goal is to foster both research and innovation in a way to address the challenges in the Big Data economy. To achieve this, specific objectives have been defined that I-BiDaaS attempts to meet. From the first two releases of the I-BiDaaS platform (MVP and 1st complete prototype), the main focus is the solution to easily be adapted by both IT and non-IT end users based on the state-of-the-art research performed in the domain of big data analytics and was reported in deliverable D1.1. Both experts and non-experts are able to carry out experiments on big data as a self-service. Diverse data analytics services are being developed based on the needs of the three different data providers and the industrial-driven and applicable in real life use cases. I-BiDaaS is able to bring together and process diverse data flows, across large, complex, and realistic inter- and intra-domains. It also provides the functionality for generation of realistic big data, to complement the real datasets in cases where real data cannot be accessed. I-BiDaaS platform, as designed successfully carries out the full cycle of the big data processing (from data collection to data visualisation). The experimentation variables that the tools and services will be validated have already been defined. Finally, I-BiDaaS is aligned with PPP activities like the Strategic Research and Innovation Agenda (BDVA SRIA), aiming to ensure the viability of the I-BiDaaS solution. I-BiDaaS architecture addresses most of the concerns identified in the European Big Data Value Strategic Research and Innovation Agenda (BDVA SRIA), in the context of the BDV reference model.

The Innovation Phase (M8-M18) of the project started right after the end of the Baseline phase (WP1) that involved the Project Set Up activity and the Positioning of I-BiDaaS. During this phase, the 1st complete prototype of the I-BiDaaS platform has been successfully delivered where all the different components have been integrated. Three different industrial sectors, i.e., finance, manufacturing and telecommunication, defined the available datasets and the use cases that the I-BiDaaS platform will be validated.

In more detail, during the 1st period, the project performed according to the planned work the following tasks:

- Definition of data assets nature and format (WP2);
- Data aggregation from completely different industrial sectors and data integration activities (WP2);
- Data extraction, transformation and data pre-processing activities (WP2);
- Test Data Fabrication (TDF) tool supports the fabrication of synthetic data sets in cases where real data are not available due to privacy restrictions (WP2);
- Design and development of interactive visualization tools for industrial self-service Big Data analytics (WP2);
- Batch processing innovative technologies for rapidly increasing historical data (WP3);
- Distributed complex event processing complemented by GPU-accelerated Analytics (WP4);

- Resource management and optimized automatic usage of computational and storage resources (WP5);
- Integration of all necessary technologies towards Big-Data-as-a-Self-Service solution (WP5);
- Definition of cross-sectoral experiment, including several data providers, in the project next phases (WP6).

The 1st complete prototype of the I-BiDaaS platform demonstrates the incorporated technologies by 4 use cases depicted in Table 4 and defined by the data providers and one generic use case. The "Analysis of relationships through IP address" use case was selected by the consortium to be the 1st use case for testing the I-BiDaaS Minimum Viable Product (MVP) that was successfully delivered by the end of the 1st year of the project (M12). The data was generated by CAIXA and IBM using the TDF tool of IBM.

I-BiDaaS achievements are closely monitored through the Key Performance Indicators (KPIs) defined by the consortium. Regarding the experimental evaluation, for each experiment, all KPIs will be monitored both in operational and technical terms. In the following Table 9, the current status of the KPIs that can be reported at this phase of the project is depicted.

Table 9: Progress with regards to I-BiDaaS KPIs

What the call states		
KPI-RI-1	Release of I-BiDaaS framework and tools under an open source non-viral license.	**3**[23]
KPI-RI-2	Increased speed of data analysis and throughput (compared to industrial-based benchmarks) by more than 10%.	**-**
KPI-RI-3	Increase in the direct access of big data analytics tools by more than 30%.	**100%**
KPI-RI-4	Define at least 2 standards related to Big Data Analytics and uptake at least 5.	**-**
KPI-RI-5	Influence at least 4 formal specifications of standards.	**-**
KPI-RI-6	Implementation of 3 data practitioners' demonstrators validating at least 80% of tools.	**4**
Impact & Exploitation KPIs		
KPI-IE-1	At least 3 I-BiDaaS tools reach market readiness level at the end of the project.	**-**
KPI-IE-2	At least 4 standalone tools and methods delivered.	**2**
KPI-IE-3	At least 6 third-party collaborations to be established for further applicability verification.	**3**
KPI-IE-4	At least 3 experiments demonstrating the tools' applicability within I-BiDaaS.	**4**
KPI-IE-5	Increased programmability for users by at least 30% compared to today, verified on at least 1 practitioner.	**-**
KPI-IE-6	Reduction of practitioners LOCs (lines of code) by 50% due to the ability to transform a sequential application into a parallel and distributed one.	**-**
KPI-IE-7	At least 1500 downloads of the tools through the project.	**-**

[23] Corresponds to the number of tools that are under an open-source non-viral license, i.e., COMPs, Hecuba, and ML pool of algorithms.

In summary, one of the goals of the I-BiDaaS consortium is to contribute to the Open Source communities to provide benefit to the European community. COMPs and Hecuba are provided under an open source license by BSC and also a pool of ML algorithms based on structured (non)convex optimization by UNSPMF that is being enriched continuously. By the end of the first period, 4 different experiments/demonstrators by 3 data providers will be implemented to validate I-BiDaaS platform and also tools' applicability. The MVP and the 1st complete prototype have been successfully delivered. Since collaboration is a fundamental concept of driving innovation, I-BiDaaS partners initiated collaborations with various third-parties such as H2020 EU projects (Toreador[24], PrEsto Cloud[25], datAcron[26]) and also with 14 different organisations that attended the I-BiDaaS Info Day – Workshop on Big Data Analytics in Novi Sad, Serbia where the MVP was demonstrated.

Finally, during the lifetime of the project, significant advances occurred in the context of data sharing. From the beginning of the project, the data providers are facing significant difficulties to share data with the consortium due to privacy restrictions they ought to comply. The TDF tool was then used to fabricate realistic data that would be as realistic as possible. During the analysis phase of the fabricated data, CAIXA realised that the insights gained was already known. To this end, CAIXA is working towards the direction of data tokenization so as to protect any sensitive information of their customers contained in the data and also to maximise the benefits gained from the project. Thus, direct access to big data analytics tools has been significantly increased. Nonetheless, the importance of the TDF tool is unquestionable since it can be used for testing purposes in cases where data cannot be made available due to data access constrains.

5.2 Involving external stakeholders in the evaluation process

Besides internal evaluation, I-BiDaaS includes several activities related with external evaluation of the project. This includes External Advisory Board (EAB) and the respective meetings, hackathons, and Info Days-Workshops organized during the course of the project. The relevant activities as well as the external feedback received are summarized below.

5.2.1 I-BiDaaS external advisory board – the introductory telcos

The External Advisory Board takes on an important role within the I-BiDaaS project. As initially reported in D6.1, the Board includes a group of experts in the field, who will further ensure the high quality and excellence of the project. The main task of the EAB is to provide external, independent analysis and recommendation on the project achievements and to bring additional competences towards a full achievement of the I-BiDaaS objectives. The EAB is composed by four (4) experts from different relevant stakeholders, industrial leaders, standardization and policy bodies, and from relevant H2020 projects. The EAB members are: **Nuria de Lama Sanchez** (female), Representative of Atos Research & Innovation to the European Commission and Vice-Secretary General of Big Data Value Association (BDVA), Spain; **George Vouros** (male), Professor, Department of Digital Systems, University of Piraeus, Greece; **Ilija Susa** (male), Co-founder of Content Insights LLC, Serbia; **Jean-Marie Hurtiger** (male), CEO of Desmond SAS Automotive Consulting,

[24] http://www.toreador-project.eu/

[25] http://prestocloud-project.eu/

[26] http://ai-group.ds.unipi.gr/datacron/

President of RENAULT light Commercial Vehicles, and former CEO of Renault Samsung Motors (Korea), France.

For the first year of the project, two EAB teleconferences were conducted, on 25th October 2018, and on 29th October 2018. The main discussions and recommendations by the EAB members were the following:

- Data integration and ingestion regarding cross-domain flow from different stakeholders.
- Management of confidentiality and protection of proprietary information along with GDPR regulation compliance and data sharing.
- Scheduling an Info Day at Novi Sad, Serbia in collaboration with Content Insights LLC to exchange ideas and discuss possible collaboration with respect to I-BiDaaS tools and technologies.

In addition, all I-BiDaaS project’s submitted deliverables became available to the EAB members.

5.2.2 I-BiDaaS external advisory board – the first physical meeting

Following the 1st EAB Introductory telco in October 2018, the 1st face-to-face I-BiDaaS EAB meeting was held on March 7th, 2019, at the premises of IBM, in Haifa, Israel. The four (4) I-BiDaaS EAB members along with representatives from all the partners of the consortium participated in the meeting. The purpose of the meeting was to present the progress that has been achieved since the introductory EAB telco and receive feedback that will guide the next steps of the project. I-BiDaaS deliverables submitted due M14 and draft versions of all presentations were made available to the EAB members beforehand aiming to receive the most valuable feedback and consultation possible.

The table below summarizes the main feedback received from the EAB members.

Table 10: Feedback received from the EAB members

Name of the EAB member	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.
Nuria de Lama Sanchez	Use BDVA Toolbox when it will be released to run I-BiDaaS benchmarks
George Vouros	Measure throughput and latency of the I-BiDaaS platform in the next release.
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.
Nuria de Lama Sanchez	Create a more realistic business model canvas in the context of I-BiDaaS Commercialization.
George Vouros	I-BiDaaS should specify the Innovations that the project will contribute to the market.
Nuria de Lama Sanchez	I-BiDaaS can contribute to the BDVA PPP newsletter.

Nuria de Lama Sanchez	I-BiDaaS should make a clearer project statement by providing a 'selling story'.
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.
Nuria de Lama Sanchez, George Vouros	I-BiDaaS should focus on Standardization activities and define the Standards that will uptake and influence.
Jean-Marie Hurtiger	Create a proforma demo to demonstrate I-BiDaaS.
Ilija Susa	I-BiDaaS should define the open-source tools and release them to receive contributions on the open source software.

5.2.3 1st I-BiDaaS Info Day-Workshop in Novi Sad

As initially reported in D6.1, the 1st I-BiDaaS Info Day-Workshop on Big Data Analytics was held on January 22nd, 2019. The participant list included the local/regional community of leading industry experts on big data analytics and data science, as well as a local network of industries interested on leveraging big data and self-service analytics in their businesses. In total, **34 persons from 14 different organisations** (12 industry companies and 2 academic organisations) participated at the event. During the workshop, I-BiDaaS experts presented the overview of the project, the architecture with its core technologies, the experimental real-life use cases from three different large industries and also to demonstrate the corresponding Minimum Viable Product.

I-BiDaaS stimulated the interest of the participants who were particularly interested in the tangible costs regarding the implementation and processes involved. In particular, the question of integration costs in corporate systems was raised alongside with questions concerning data management costs. All the questions above made it clear that there is a strong interest in the commercial use of both the product and the technologies adopted. I-BiDaaS team members clarified that the conclusions regarding the effectiveness of the project as a self-service solution for Big Data Analytics in business domains are exported from the MVP. Thus, the development of a business plan will provide a better understanding of the flexibility and implementation of the final product. The other important point is related to data pre-processing and transformation that occurs before actual processing. Since data preparation is a fundamental stage of data analysis and also an ongoing process, participants highlighted the need of incorporating interactive, self-service visualisation tools into the I-BiDaaS platform. Challenges regarding both the real-time data management & transformation and a friendly UI of the final product were also pointed out by many participants. To this end, I-BiDaaS experts requested active involvement on the project and acknowledged the need for feedback from stakeholders after testing the product. Thus, the participants were invited to take part in future activities that will be organised (i.e., hackathons, workshops, etc.)

5.2.4 CRF's hackathon at Campus Melfi

As indicated in D6.1, CRF organised a hackathon in Campus Melfi (Italy) from 18th to 19th June, 2019 in order to share knowledge on the automated data processing for manufacturing and to start defining a business case for the big data analysis across different processes and plants. The aim of the hackathon is on one side to develop methodologies and tools for the predictive analysis of massive manufacturing data, showing heterogeneous real-time acquisition, massive industrial data management and intelligent decision-making, and on the other side to define the metrics, objectives, AS-IS and TO-BE situations for the implementation of big data analysis in automotive industrial processes.

In the CRF's Hackathon students, PhDs, researchers and employees from SMEs were asked to develop an innovative solution for the analysis, visualisation and architecture of complex data collected from a die casting process. The dataset was previously processed by the CRF which structured and anonymised the large volume of Aluminium Casting data.

CRF experts explained to the participants the relevant information concerning the die casting process, from which the dataset provided was extracted, illustrating the different parameters for the analysis. Six teams were formed after interviewing the participants in order to understand their skills on Analysis, Visualization and Architecture / Hardware, in order to create heterogeneous groups able to perform all the different actions required.

To motivate the teams, improve communication and collaboration, a team building was organised, during which the individual participants were involved in two role-playing games. In the first one, participants were asked to make a chain, as long as possible, with four sheets available. Later they were given time to confront each other in order to do again the same game with common experience, but with only two sheets. In the second game, participants were asked to form a circle. So everyone was asked to sit on the legs of the person behind at the same time. Once the collaborative spirit was created, the groups were accommodated in different work areas. At the same time the I-BiDaaS project partners were able to analyse the case study, even if out of competition.

The teams worked by using different software (Minitab, Excel, R) and methods: Neural Networks (Multi Layer Perceptron MLP), Decision Trees, Random Forest and Generalized Additive Model (GAM) for Big Data.

Each group was able to find valuable results using also different techniques than the ones used in the analysis performed by CRF previously.

In particular, this is the summary from a technical point of view:

- **Benchmark:** With the same data set, the I-BiDaaS team achieved 80% classification accuracy, while best reported by other teams was 73-74%.
- **Validation:** Other teams found similar features as the I-BiDaaS team did, while the modelling approach was different. This cross-validates the project approach with others.
- **Casting process phases:** Other teams took into account the three phases of the casting process while the project team did not.
- **Data cleaning:** Other teams pointed to a problem in Result_2 data, i.e., there was class label 4, while we did not take this into account.

Other important outcomes/advances include:

- **Break silos:** I-BidaaS got access to real anonymized data and engine casting process, the data is at least order of magnitude larger than the data previously available to the consortium
- **Data fabrication:** New approaches for fabricating data *with labels* were brainstormed and delivered during the hackathon.

In synthesis, as a result of the hackathon, we produced a new random forest model and a new neural network model (classifier) for the CRF foundry use case. As mentioned before, our random forest model had 80% accuracy versus 74% accuracy achieved by other hackathon teams.

CRF is exploiting the results obtained during the hackathon to address the opportunities and challenges of industrial big data analytics produced by diverse sources in manufacturing spaces, such as sensors, devices and humans. The project and hackathon results will be evaluated based on their effective capability to increase the production efficiency and to reduce energy consumption and cost of production.

5.2.5 Hackathon at TID

As initially reported in D6.1, TID plans to organise a hackathon at its premises in Barcelona, where participants (startups, SMEs, etc.) can collaborate on creating new innovative ideas for using data offered by TID. More specifically, TID is considering making available (to the participants of the hackathon) data from Movistar+ (a Video on Demand service) and BotCorn (a recommendation service for movies/tv series). The aim of the hackathon is to help TID break internal and external data silos in a secure, privacy-preserving manner that respects any applicable legal requirements. The results of the hackathon will be evaluated based on the creativity, innovation, and business value of the produced solutions, but also according to the level exposure of the project's results to relevant communities, as well as the visibility and public awareness achieved.

6 Conclusion

This deliverable reports on the progress of the specification of the operational experiments and initiates the experimentation phase of the I-BiDaaS project (M18-M30). Special effort has been made to detail all experiments, whilst four experiments have been fully considered according to the I-BiDaaS plan, in terms of: the dataset preparation; experimental workflow specification and participants' identification; and the definition of the experimental evaluation indicators according to the business objectives.

These experiments will be performed using the first release of the prototype. The initial results will be reported in the first evaluation report (D6.3 due M22). The specification and refinement of the operational experiments will continue during the experimentation phase of the project based on data availability and released functionalities of the I-BiDaaS platform.

To ensure a greater objectivity and validity of experimental results I-BiDaaS also considers the participation of external stakeholders, both domain experts and prospective users. To this end the deliverable reports on a number of activities aiming to the involvement of external stakeholders as well as on the definition of a cross-sectoral experiment.

7 References

[1] **I-BiDaaS project 2018a.** Deliverable D1.3: Positioning of I-BiDaaS, Public Report, available at https://www.ibidaas.eu/deliverables

[2] **I-BiDaaS project 2018b.** Deliverable D2.1 Data assets and formats, Confidential Report

[3] I-BiDaaS project 2019a. Deliverable D2.4 Universal Messaging Bus (interim version), Public Report, available at https://www.ibidaas.eu/deliverables

[4] **I-BiDaaS project 2019a.** Deliverable D3.2 Batch Processing Analytics module implementation as part of I-BiDaaS solution, Public Report, available at https://www.ibidaas.eu/deliverables

[5] **I-BiDaaS project 2018c.** Deliverable D5.2 Big-Data-as-a-Self-Service Test and Integration Report (first version), available at https://www.ibidaas.eu/deliverables

[6] **I-BiDaaS project 2019b.** Deliverable D6.1 Evaluation report (interim version), Confidential Report

Annex A Quantitative Evaluation Questionnaire

The checklist list that follows refers to the first phase and should be completed by all technology providers. It contains a number of indicators (defined during the scoping and planning phases of the experimental process, reported in D1.3) that should be measured in order to quantitatively evaluate the I-BiDaaS platform.

Please check that the indicators are accurate and complete (making all necessary changes) and indicate the metrics, baseline values and/or benchmark to be used when assessing the quality of the I-BiDaaS platform.

Technical evaluation checklist

	Indicator	**Metric**	**Baseline value / Benchmark**[27]	**measured I-BiDaaS value**[28]
MVP level	Usability			
	Scalability			
	Operational Performance			
	Availability			
	Reliability			
	Data Security			
	Privacy			
	Compliance			
	Cost			
Test Data Fabrication (IBM)	Scalability			
	Validity			
	Performance			
	Compliance			
	Availability			
Universal Messaging (SAG)	Scalability			
	Operational performance			
	Availability			
	Reliability			
	Data Security			
	Compliance			

[27] To be completed by technology providers

[28] To be recorded during the evaluation

	Indicator	Metric	Baseline value / Benchmark[27]	measured I-BiDaaS value[28]
Advanced ML (UNSPMF)	Quality of results			
CEP Engine (SAG)	Scalability			
	Availability			
	Reliability			
GPU-accelerated pattern matching (FORTH)	Performance			
Hecuba DBS (BSC)	Scalability			
	Operational Performance			
	Availability			
	Reliability			
Qbeast (BSC)	Scalability			
	Operational performance			
	Availability			
	Reliability			
Resource management and orchestration module (ATOS)	Operational performance			
	Scalability			
	Availability			
	Reliability			
Visualisation Tool (AEGIS)	Operational performance			
	Scalability			
	Availability			
	Reliability			
	Usability (Efficiency)			
	Usability (Satisfaction)			

Annex B Qualitative Evaluation Questionnaire

The following checklist list refers to the second phase and should be completed by use case providers (sections 1, 2 and 3) and technology providers (sections 3 and 4).

It aims to refine the business indicators and associated metrics (defined during the scoping and planning phases of the experimental process, reported in D1.3) that should be measured in order to qualitatively evaluate the I-BiDaaS platform.

Please check that the indicators and the metrics are accurate and complete (making all necessary changes) and indicate the baseline values and/or benchmark to be used when assessing the quality of the I-BiDaaS platform.

It should be noted that one checklist should be completed for each use case (and associated experiments).

1. Use case overview

No	
Name	
Provider	
Description	
Business Goal	

2. Business Indicators

	Indicator	**Metric**	**Baseline value / Benchmark**	**measured I-BiDaaS value**
Business KPIs				
Application Level Performance indicators[29]				
Platform level performance indicators[30]				

[29] Related to specific I-BiDaaS component

[30] Related to the integrated I-BiDaaS platform

3. Experiment overview[31]

Experiment Name		
Experiment Description (workflow)[32]		
Participants	Quality assurance and control managers	
Please indicate number	Data analysts	
in each category	Financial administrators	
	Infrastructure engineers	
	IT security personnel	
Processing Type[33]		

[31] Repeat this section if more there are more than one experiments

[32] Describe the steps that should be performed by the platform users

[33] Batch / Streaming