


Tailored IoT & BigData Sandboxes and Testbeds for Smart,  
Autonomous and Personalized Services in the European  
Finance and Insurance Services Ecosystem



## D7.1 – Report on Pilot Sites Preparation - I

Revision Number	3.0
Task Reference	T7.1
Lead Beneficiary	ATOS
Responsible	Jose Gato Luis
Partners	ATOS, GFT, IBM, FTS, HPE, SILO, ENG, ISPRINT, SIA, LIB, NBG, AKTIF, BOS, PI, BPFI, DYN, GEN, JRC, PRIVE, CP, WEA, RB, LXS, UNP, RRD, FBK, NUIG, NOVA, BOUN, JSI, ORT, CTAG, GRAD, ABILAB, BOI, AGRO, BANKIA, BOC, UPRC
Deliverable Type	R
Dissemination Level	PU
Due Date	2020-11-30
Delivered Date	2020-12-14
Internal Reviewers	GFT, INNOV
Quality Assurance	INNOV
Acceptance	WP Leader Accepted and Coordinator Accepted
EC Project Officer	Pierre-Paul Sondag
Programme	HORIZON 2020 - ICT-11-2018
	This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement no 856632

## Contributing Partners

Partner Acronym	Role <sup>1</sup>	Author(s) <sup>2</sup>
ATOS	Deliverable leader	Jose Gato Luis, Ignacio EliceGUI Maestro (ATOS)
GFT	Cluster Leader	Vittorio Monferrino, Maurizio Megliola (GFT)
CrowdPolicy	Cluster Leader	Marinos Xynarianos (CP)
WeAnalyze	Cluster Leader	Carlos Albo (WeAnalyze)
Cluster 1	Pilot Leaders and Technical Proxies	Barbara Cacciamani, Marco Rotoloni (ABILAB) Javier Sancher (bankia), antoni munar (gft)  Petra Ristau (jrc), John Soldatos (Innov), Nikos Kapsoulis (INNOV), Despoina Kyriazis (INNOV)
Cluster 2	Pilot Leaders and Technical Proxies	Richard Walsh (NUIG), Martin Serrano (INSIGHT) Pablo Carballo (PRIVE) John Kaldis (RB) Silvio Walser (BOC) Dimitris Kotios (UPRC) Manolis Syllignakis, Nikos Droukas, George Kanellis (NBG)
Cluster 3	Pilot Leaders and Technical Proxies	Maja Skrjanc (JSI) Sabina Podkriznik (BOS)Can Ozturan (BOUN)Susanna Bonura (ENG) Massimiliano Aschi (PI)
Cluster 4	Pilot Leaders and Technical Proxies	Jose Gato Luis, Ignacio EliceGUI Maestro (ATOS) Aristodemos Pnevmatikakis (ISPRINT)
Cluster 5	Pilot Leaders and Technical Proxies	Carlos Albo (WeAnalyze) Gregory Mygdakos (AGRO) Lukas Linden (GEN)

<sup>1</sup> Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

<sup>2</sup> Can be left void

## Revision History

Version	Date	Partner(s)	Description
0.1	2020-09-21	ATOS	ToC Version
0.2	2020-10-06	ATOS	Executive Summary, Introduction
0.3	2020-10-06	ATOS	Pilot 11
0.4	2020-11-06	ATOS+ClusterLeader+Pilots	Pilot 6, 7, 9, 10, 12
0.5	2020-11-11	ATOS+ClusterLeader+Pilots	Pilot 4, 5, 8, 13
0.6	2020-11-16	ATOS+ClusterLeader+Pilots	Pilot 1,2, 14, 15
0.7	2020-11-17	ATOS	Version ready for internal review
0.8	2020-11-25	ATOS+GFT+INNOV	Internal review changes
0.9	2020-11-25	Cluster Leaders	Internal review changes
0.10	2020-12-2		Ready for a new review
0.11	2020-12-9	ATOS+INNOV	Second Internal review changes. Ready for Q&A
2.0	2020-12-14	INNOV	Version for Quality Assurance
3.0	2020-12-14	GFT Accepted	Version for submission

## Executive Summary

This deliverable aims to specify different aspects of each large-scale pilot: readiness; development; and validation of different services and components. Validation is a core pillar, due to one of the main objectives of INFINITECH is to test innovative (IoT, BigData, AI, ML, Blockchain and more) technologies towards improve business services in the Financial and Insurance sector. Specifically, the present deliverable reports on the readiness of the various pilot sites to test the INFINITECH innovative AI, IoT and BigData technologies into the testbeds/sandboxes that are developed during the project, while validating their ability to improve the business processes of end-user organizations (i.e. financial organizations, banks, and FinTech firms)

In summary, this deliverable reports for each one of the pilots sites the following information:

- A General overview of the status of the pilot, including its main business and technical objectives
- The development status of the different components and services that comprise each pilot system.
- The status of the integration of a subset of their components as part of a Proof-of-Concept (PoC) pilot system.
- Information on the availability and deployment status of the testbed/sandbox, where the pilot' final infrastructure will be deployed and validated.

Pilots have already contributed to other previous deliverables and tasks (requirements, user stories, security, policies, technologies, services, RA, etc). Therefore, this deliverable builds on top of these contributions. However, it also integrates and extends them, through illustrating how individual technical activities are enabling the integration and deployment of a complete pilot system with relevance for the end-users (i.e. financial organizations, banks, FinTechs). Overall, the present deliverable focuses not only on individual contributions, but rather on the overall readiness of the pilot sites and the pilots' frameworks as a whole.

Note that the deliverable does not elaborate on the following aspects that have been already covered in earlier deliverables (notably WP2 and WP5 deliverables):

- The deliverable does not describe user stories, requirements, business services, etc. However, it provides a high-level overview of them to facilitate a better understanding of the pilots.
- The deliverable does not go in detail about data sources and data formats. However, it outlines information about existing/collected data, size, availability, and plans for future data capture.
- The deliverable does not describe in detail the components and services that comprise the pilot systems. Nevertheless, it outlines the current development status of the various components, which is key for understanding the readiness of the pilot sites.

**In summary: The document presents readiness of each pilot according to its objectives.**

As already outlined the deliverable presents the status of an initial PoC implementation for each pilot. This PoC enables a first demonstration of the viability and applicability of the various INFINITECH technologies that support the pilots. The various pilots PoCs demonstrate the different developments accomplished up to date and serve as a basis for ensuring that the pilots' developments are on the right track, while identifying points that need attention where required.

The current status of testbeds/sandboxes is also an important part of the deliverable, because it directly affects the readiness and demonstrability of each pilot. Therefore, a quick overview of each testbed/sandbox is covered by each pilot.

In a future second version (update) of this deliverable, it will be possible to go in depth about more technical details and specifications of each pilot. It is being prepared a source code repository with CI/CD tools that will provide these details in the next update.

## Table of Contents

1	Introduction.....	12
1.1	Objective of the Deliverable.....	14
1.2	Insights from other Tasks and Deliverables.....	15
1.3	Structure.....	15
2	Pilot sites preparation .....	17
2.1	Pilot #1 Invoices Processing Platform for a more Sustainable Banking Industry .....	17
2.1.1	Technological components and Services.....	18
2.1.2	Data sets status .....	18
2.1.3	Testbed .....	19
2.1.4	Others non-technical requirements .....	19
2.1.5	Implementation of a first Proof of Concept .....	19
2.1.6	Next steps and timeline.....	20
2.1.7	Conclusions - Issues and Barriers .....	21
2.2	Pilot #2 Real-time risk assessment in Investment Banking .....	21
2.2.1	Technological components and Services.....	22
2.2.2	Data sets status .....	23
2.2.3	Testbed .....	24
2.2.4	Others non-technical requirements .....	25
2.2.5	Implementation of a first Proof of Concept .....	25
2.2.6	Next steps and timeline.....	27
2.2.7	Conclusions - Issues and Barriers .....	27
2.3	Pilot #3 Collaborative Customer-centric Data Analytics for Financial Services .....	28
2.4	Pilot #4 Personalised Portfolio Management.....	28
2.4.1	Technological components and Services.....	29
2.4.2	Data sets status .....	30
2.4.3	Testbed .....	32
2.4.4	Others non-technical requirements .....	33
2.4.5	Implementation of a first Proof of Concept .....	33
2.4.6	Next steps and timeline.....	34
2.4.7	Conclusions - Issues and Barriers .....	34
2.5	Pilot #5b Business Financial Management (BFM) tools delivering a Smart Business Advise.....	35
2.5.1	Technological components and Services.....	36
2.5.2	Data sets status .....	37
2.5.3	Testbed .....	38
2.5.4	Other non-technical requirements.....	38

2.5.5	Implementation of a first Proof of Concept .....	38
2.5.6	Next steps and timeline .....	39
2.5.7	Conclusions - Issues and Barriers .....	39
2.6	Pilot #6 Personalized Closed-Loop Investment Portfolio Management for Retail Customers.....	40
2.6.1	Technological components and Services.....	41
2.6.2	Data sets status .....	42
2.6.3	Testbed .....	43
2.6.4	Other non-technical requirements.....	44
2.6.5	Implementation of a first Proof of Concept .....	44
2.6.6	Next steps and timeline .....	45
2.6.7	Conclusions - Issues and barriers.....	45
2.7	Pilot #7 Operation Whitetail - Avoiding Financial Crime .....	45
2.8	Pilot #8 Platform for AML supervision .....	46
2.8.1	Technological components and Services.....	47
2.8.2	Data sets status .....	48
2.8.3	Testbed .....	49
2.8.4	Others non-technical requirements .....	49
2.8.5	Implementation of a first Proof of Concept .....	50
2.8.6	Next steps and timeline.....	52
2.8.7	Conclusions - Issues and Barriers .....	52
2.9	Pilot #9 Analysing Blockchain Transaction Graphs for Fraudulent Activities .....	52
2.9.1	Data sets status .....	54
2.9.2	Testbed .....	55
2.9.3	Other non-technical requirements.....	55
2.9.4	Implementation of a first Proof of Concept .....	55
2.9.5	Next steps and timeline.....	56
2.9.6	Conclusions - Issues and Barriers .....	57
2.10	Pilot #10 Real-time cybersecurity analytics on financial transactions’ data .....	57
2.10.1	Data sets status .....	61
2.10.2	Testbed .....	61
2.10.3	Next steps and timeline.....	62
2.10.4	Implementation of a first Proof of Concept .....	63
2.10.5	Conclusions - Issues and Barriers .....	64
2.11	Pilot #11. Personalized insurance products based on IoT connected vehicles. ....	65
2.11.1	Technological components and Services.....	67
2.11.2	Data sets status .....	68
2.11.3	Testbed .....	69
2.11.4	Others non-technical requirements .....	69
2.11.5	Implementation of a first Proof of Concept .....	70

2.11.6	Next steps and timeline.....	70
2.11.7	Conclusions - Issues and Barriers .....	71
2.12	Pilot #12 Real world data for novel health insurance products.....	71
2.12.1	Technological components and Services.....	73
2.12.2	Data sets status .....	73
2.12.3	Testbed .....	74
2.12.4	Other non-technical requirements.....	75
2.12.5	Implementation of a first Proof of Concept .....	75
2.12.6	Next steps and timeline.....	75
2.12.7	Conclusions - Issues and Barriers .....	76
2.13	Pilot #13 Alternative and automated insurance risk selection and insurance product recommendation for SME's.....	76
2.13.1	Technological components and Services.....	79
2.13.2	Data sets status .....	79
2.13.3	Testbed .....	79
2.13.4	Other non-technical requirements.....	80
2.13.5	Implementation of a first Proof of Concept .....	80
2.13.6	Next steps and timeline.....	81
2.13.7	Conclusions - Issues and Barriers .....	81
2.14	Pilot #14 Big Data and IoT for the Agricultural Insurance Industry.....	82
2.14.1	Technological components and Services.....	83
2.14.2	Data sets status .....	84
2.14.3	Testbed .....	85
2.14.4	Others non-technical requirements. ....	85
2.14.5	Implementation of a first Proof of Concept .....	86
2.14.6	Next steps and timeline.....	87
2.14.7	Conclusions - Issues and Barriers .....	87
2.15	Pilot #15 Inter-Banking Open Pilot .....	88
2.15.1	Technological components and Services.....	89
2.15.2	Data sets status .....	89
2.15.3	Testbed .....	90
2.15.4	Others non-technical requirements .....	90
2.15.5	Implementation of a first Proof of Concept .....	90
2.15.6	Next steps and timeline.....	90
2.15.7	Conclusions - Issues and Barriers .....	90
3	Conclusions.....	91

## List of Figures

Figure 1: Pilot#1 Reference Architecture .....	17
Figure 2: Pilot #1 Main Components Interactions.....	20
Figure 3: Pilot #1 Timeline .....	20
Figure 4: Pilot #2 Data Science Pipeline .....	22
Figure 5: Pilot #2 Deployment Diagram .....	23
Figure 6: Pilot #2 Dashboard for parameter configuration and visualization .....	26
Figure 7: Pilot #2 Charts for Three VaR Calculation Method.....	26
Figure 8: Pilot #4 Roles and Services .....	28
Figure 9: Pilot #4 Reference Architecture (From D2.13) .....	29
Figure 10: Pilot #4 Process flow diagram .....	30
Figure 11: Pilot #4 High-Level Architecture .....	32
Figure 12: Pilot #4 Components .....	34
Figure 13: Pilot #5b Reference Architecture .....	36
Figure 14: Pilot #5 Services.....	37
Figure 15: Pilot #6 Personalized Closed-Loop Investment Portfolio Management for Retail Customers .....	40
Figure 16: Pilot #6 Reference Architecture (From D2.13) .....	41
Figure 17: Pilot #6 High Level Architecture (updated From D6.1) .....	42
Figure 18: Pilot #6 Data collection PoC architecture.....	44
Figure 19: Pilot #7 RA (From D2.13) .....	46
Figure 20: Pilot #8 Reference Architecture (From D2.13) .....	48
Figure 21: Pilot #8 Risk assessment tool data flow .....	50
Figure 22: Sector risk Assessment View .....	50
Figure 23: Sector Risk Assessment view- historical view .....	51
Figure 24: Inherent risk and control environment view.....	51
Figure 25: Bank Profile View.....	51
Figure 26: Pilot #8 Timeline to implement user stories and components .....	52
Figure 27: Pilot #9 Reference Architecture .....	53
Figure 28: Pilot #9 Architecture of the Proof of Concept.....	55
Figure 29: Pilot #10 Reference Architecture .....	58
Figure 30: Pilot #10 Testbed logic .....	62
Figure 31: Pilot #10 PoC (October 2020) .....	63
Figure 32: Pilot #10 Clustering workflow .....	63
Figure 33: Pilot #10 Clustering Results.....	64
Figure 34: Pilot #11 Personalized insurance products based on IoT connected vehicles overview .....	65
Figure 35: Pilot #11 Reference Architecture (From D2.13) .....	66
Figure 36: Pilot #11 High Level Architecture (updated from INFINITECH D6.1).....	68
Figure 37: Pilot #11 Data collection PoC architecture.....	70
Figure 38: Pilot #12 Real world data for novel health insurance products overview .....	71
Figure 39: Pilot #12 Reference Architecture (From D2.13) .....	72
Figure 40: Pilot #12 High Level Architecture (updated From D6.1) .....	72
Figure 41: Pilot #12 PoC user participation. Users’ demographics (left) and a particular user’s steps data (right).....	74
Figure 42: Pilot #13 summary.....	77
Figure 43: Pilot #13 Reference Architecture (from D2.13).....	77
Figure 44: Pilot #13 Customer Acquisition Funnel .....	78
Figure 45: Pilot #13 Platform schema .....	78
Figure 46: Pilot #13 Implemented Proof of Concept.....	81
Figure 47: Pilot #14 Reference Architecture .....	82
Figure 48: Pilot #14 Data sources .....	84



Figure 49: Pilot #14 testbed .....	85
Figure 50: Pilot #14 Proof of Concept .....	86
Figure 51: Pilot #14 Data requirements for developing a pricing framework.....	87
Figure 52: Pilot #15 Steps and main objectives.....	89

## List of Tables

Table 1 Pilots about Smart, Reliable and, Accurate Risk and Scoring Assessment .....	12
Table 2 Pilots about Personalized Retail and Investment Banking Services .....	12
Table 3 Pilots about Financial Crime and Fraud Detection .....	13
Table 4 Personalized Usage Based Insurance Products .....	13
Table 5 Configurable and Personalized Insurance Products .....	13
Table 6 Pilot #1 Data Sets Status .....	19
Table 7 Pilot #2 Data Sets Details .....	24
Table 8 Pilot #4 Data Sets Summary.....	31
Table 9 Pilot #6 Hardware/Software Requirements for the Testbed.....	43
Table 10: Testbed Specification for Pilot #8.....	49
Table 11 Ethereum Mainnet Blockchain Dataset Details .....	54
Table 12 Bitcoin Blockchain Dataset Details .....	55
Table 13 Pilot #10 Components Readiness .....	61
Table 14 Pilot’s #11 hardware testbed (first analysis). From D6.1.....	69
Table 15 Pilot’s #12 hardware testbed.....	74
Table 16 Pilot #13 Data Sets Details .....	79
Table 17 Pilot #13 Hardware/Software Requirements .....	80
Table 18 High Level Overview of Pilots’ Implementation Status .....	92

## Abbreviations/Acronyms

### Abbreviation Definition

AgI	Agricultural Insurance
AI	Artificial Intelligence
AIGO	AI Based Portfolio Optimization Process
AML	Anti-Money Laundering
AoI	Area of Interest
API	Application Programming Interface
AWS	Amazon Web Services
AWS EC2	Amazon Web Services Elastic Compute Cloud
BDA	Big Data Analytics
BDVA	Big Data Value Association
BFM	Business Financial Management
CSV	Comma Separated Value
BD	Database
DL	Deep Learning
DNS	Domain Name Service
DWH	Data Warehouse
EO	Earth Observation
ERC20	Ethereum Requests for Comment
FIBO	Financial Business Ontology
FX	FOREX
GB	GigaBytes
HDFS	Hadoop Distributed File System
HPC	High Performance Computing
IDM	Identity Management
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
KYC	Know Your Customer
MB	Message Broker
ML	Machine Learning
MPI	Message Passing Interface
MS	Microsoft
NLP	Natural Language Processing

D7.1 – Report on Pilot Sites Preparation - I

NN	Neural Network
OCR	Optical Character Recognition
PAMLS	Platform for AML
PCTU	Top-ups phone credit
PDF	Portable Document Format
PoC	Proof of Concept
RA	Reference Architecture
ReM	Resource Management
RoM	Role Management
RWD	Real World Data
SEPA	Single Euro Payments Area
SFTP	Secure File Transfer Protocol
STFTS	Secure telematic fund transmission system
SHARP	Smart, Holistic, Autonomy, Personalized and Regulatory Compliant
SME	Small Medium Enterprise
SMWCA	Sending Money to people Who do not have a Current Account
TRL	Technology Readiness Level
TRY	Turkish Lyra
UAT	User Acceptance Testing
UI	User Interface
VAR	Value at Risk

# 1 Introduction

The main objective of WP7 is to implement the different pilots of INFINITECH, through:

- Developing and integrating the INFINITECH pilot services, notably services that feature the SHARP (Smart, Holistic, Autonomy, Personalized and Regulatory Compliance) properties.
- Deploying and testing the services in testbeds/sandboxes and
- Validating the services against the business requirements of the pilots towards enhancing/improving business services of the end-users.

According to the DoA, WP7 aims at:

- integrating and deploying SHARP (Smart, Holistic, Autonomy, Personalized and Regulatory Compliant) services that cover all aspects of BigData, AI and IoT applications in the financial and insurance sectors.
- validating the services from both, technical and business viewpoints.
- receiving and analysing stakeholders' feedback, while using this feedback to improve technologies, sandboxes and business models.

This is the first deliverable of WP7 which focuses on the readiness of the different aspects and status of each of the current 15 pilots of the project. These pilots are grouped in five thematic clusters and are assigned different numbers/codes as follows:

<b>Cluster 1: Smart and Reliable Scoring, Risk and Service Assessment</b>	
<b>Pilot #1</b>	<b>Invoices Processing Platform for a more Sustainable Banking Industry</b> (BANKIA, GFT, FBK, RB, INSO)
<b>Pilot #2</b>	<b>Real-time risk assessment in Investment Banking</b> (JRC, INNOV, GFT)
<b>Pilot #15</b>	<b>Open Inter-Banking Pilot</b> (ABILAB, GFT, HPE)

Table 1 Pilots about Smart, Reliable and, Accurate Risk and Scoring Assessment

<b>Cluster 2: Personalized Retail and Investment Banking Services</b>	
<b>Pilot #3</b>	<b>Collaborative Customer-centric Data Analytics for Financial Services</b> (BPFI, NUIG, Selected Bank)
<b>Pilot #4</b>	<b>Personalised Portfolio Management (“Why Private Banking cannot be for everyone?”)</b> (PRIVE, RB)
<b>Pilot #5b</b>	<b>Business Financial Management (BFM) tools delivering a Smart Business Advise</b> (BOC, UPRC, GFT, CP)
<b>Pilot #6</b>	<b>Personalized Closed-Loop Investment Portfolio Management for Retail Customers</b> (NBS, CP, RB, PRIVE)

Table 2 Pilots about Personalized Retail and Investment Banking Services

<b>Cluster 3: Financial Crime and Fraud Detection</b>	
<b>Pilot #7</b>	<b>Avoiding Financial Crimes</b> (CXB, FTS, GFT, FBK, INSO)
<b>Pilot #8</b>	<b>Platform for AML supervision (PAMLS)</b> (BOS, JSI)
<b>Pilot #9</b>	<b>Analysing Blockchain Transaction Graphs for Fraudulent Activities</b> (AKTIF, BOUN)

<b>Pilot #10</b>	<b>Real-time cybersecurity analytics on financial transactions' data</b> (PI, ENG)
------------------	---

Table 3 Pilots about Financial Crime and Fraud Detection

<b>Cluster 4: Personalized Usage Based Insurance Products</b>	
<b>Pilot #11</b>	<b>Personalized insurance products based on IoT connected vehicles</b> (ATOS, CTAG , GRAD, DYN)
<b>Pilot #12</b>	<b>Real World Data for Novel Health-Insurance products</b> (ISPRINT, RRD, SILO, DYN, GRAD)

Table 4 Personalized Usage Based Insurance Products

<b>Cluster 5: Configurable and Personalized Insurance Products</b>	
<b>Pilot #13</b>	<b>Alternative and automated insurance risk selection and insurance product recommendation for SME's (WEA, LXS)</b>
<b>Pilot #14</b>	<b>Big Data and IoT for the Agricultural Insurance Products Industry</b> (GEN, AGRO)

Table 5 Configurable and Personalized Insurance Products

While the present deliverable reports on the status and readiness of all pilots, there are pilot that are more advanced than others, while some pilots have only preliminary progress. The latter is a result of the fact that few pilot partners joined with the project late, as part of recent amendments of the INFINITECH Contract. Overall, when reading the pilot reports the following information should be taken into account:

- Pilot #3, Collaborative Customer-centric Data Analytics for Financial Services, has been affected from some internal changes in their partners. Its scope has been recently redefined and hence its implementation is in its early stage. This was mainly due to the late inclusion of the pilot business owner in the project.
- The business owner of Pilot #7, Advanced Financial Crime Risk Model and Scoring, (Caixabank) has recently joined the project and hence the pilot specifications are still in their early stages. Likewise, the implementation progress of the pilot lags behind other pilots.
- Pilot #15, Open Inter-Banking Pilot, joined the project in a later stage. It is a pilot based on Open Calls and the engagement of Banks outside the consortium. The pilot has completed the open call phase and hence its implementation status is at an early stage. The scope and ambition of the pilot increased its complexity as well, having as objective to solve a shared problem among different banks and financial institutions, to craft a solution following a collaborative and pre-competitive approach.

Other special clarification about Pilot #5:

- Pilot #5, Business Financial Management (BFM) tools delivering a Smart Business Advise, had originally two complementary versions. With Liberbank and Bank of Chyprus leading each one of these versions (5a more oriented to B2C and 5b to B2B). After Liberbank withdrawal from INFINITECH, the Pilot is now active as the Pilot #5b, and it has not been affected, proceeding smoothly and as planned. The work done by Liberbank has been and will be exploited by Pilot#5b due to their very similar requirements and objectives.

INFINITECH follows a result-oriented approach to manage the 15 pilots of the project. More specifically, WP7 has as main objective the overall coordination of the pilot, both inside and outside the specific work package. Indeed, several activities have to be performed to work and collaborate along the other WPs, starting with WP2 to define requirements and specifications, then with WP3, WP4 and WP5 to align the development of

technological components, WP6 to structure the testbeds and to deploy the sandbox instances, WP8 to feed the marketplace and platform, and finally with WP9 to disseminate and exploit the results.

The high-level organization introduced in the WP7 to achieve and successfully manage the communication inside out the project is mainly defined as follow:

- Each pilot is contained into a cluster of pilots with similar objectives and thematic relevance. This is performed to facilitate collaboration and experience sharing, find synergies among pilots, and share know-how and methodologies to achieve results reducing effort.
- Each of these clusters belongs to one task in WP7. So, task leaders (T7.2, T7.3, T7.4, T7.5 and T7.6) are a synonymous of cluster leaders. These leaders have been reconfigured (during amendment) to position partners with enough experience in managing EU projects and with technical and business knowledge, to ensure smooth procedures and results achievement.
- Internal bi-weekly meetings between WPL (ATOS) and Cluster Leaders, analysing results, risk management, critical paths, possible barriers and delays, and mitigation strategies. Different cluster meetings organized on demand, complemented with specific pilot's meetings when needed.
- Contributions from pilots to deliverables in other WPs are centralized and managed by WP7 leader and tasks leaders, leveraging on the direct contact among the actors, funnelling the requested contribution on a reverse approach, supporting pilots in their delivery of results and then concentrating them to produce proof of work (deliverables).
- Structured and frequent communication with Pilots to check and assess their readiness and their alignment with INFINITECH and innovation.
- Use of Common templates for reporting, contributing to other deliverables and pilot analysis.

## 1.1 Objective of the Deliverable

The main objective for this deliverable aims to show the readiness of each pilot about implementing/developing/deploying its different services/components, with the final objective to provide enhanced/new business services. In particular:

- General overview and status of the pilot.
- Different components and services (technical development status)
- Description and objectives for a first PoC ready by the release of the current deliverable
- Testbed/sandbox availability

During the first year of the project, WP7 pilots have been actively contributing to different WPs and Deliverables. The intention of this deliverable is not to repeat those contributions again, but pointing them quickly for easy understanding of each pilot. This deliverable focuses on pilots' readiness: current available components developed, data already gathered and available, demonstrations of different technologies, etc.

In order to start validating pilots and technologies, each pilot has defined a first Proof of Concept (PoC). This PoC shows the status of the technologies and demonstrates some of the objectives and future business services. It is important to remark that the objectives and the scope of this first PoC have been defined individually by each Pilot Leader.

Next version of this deliverable will focus more in detail about each developed service, feedback and technologies' testing.

## 1.2 Insights from other Tasks and Deliverables

As it has been pointed out above, WP7 has been contributing to different WPs and deliverables (through the pilot's management methodology).

Following list shows the more relevant contributions from WP7, mainly to WP2, to set requirements, services, user stores, etc:

- D2.1 User Stories and Stakeholders' Requirements -I
  - All the pilots filled a form to capture user stories and requirements in the way of Jobs To be Done, Gains, Pains, Roles, etc
- D2.3 Reference Scenarios and Use Cases – Version I
  - All the pilots contributed with business services and the functional services
- D2.5 – Specifications of INFINITECH Technologies - I
  - All the pilots filled a form detailing the technologies to be developed to implement the user stories.
- D2.7 Security and Regulatory Compliance Specifications – I
  - How regulations affect INFINITECH projects with regards to the pilots use cases.
  - Awareness of the pilots regarding personal data and the details on how to deal with the various requirements resulting from the GDPR.
- D2.9 – Initial Specification of Testbeds, Data Assets and APIs - I
  - A first quick look to the data sources to be deployed into the testbeds
- D2.11 – Data Models Specification - I
  - A first quick look to the data models for each pilot
- D2.13 – INFINITECH Reference Architecture - I
  - This is a crucial deliverable in the project objectives. An intensive work was made here to define a generic Reference Architecture (Based on the BDVA RA) that all pilots can fit in.
  - Each pilot was defined about this RA using the components pointed out in the D2.5. Fitting those technologies into the different layers of the RA
  - The business services defined in D2.3 were also explained as a pipeline with the different components of the RA.
  - Pilots agreed to follow this architecture and the concepts of “micro-services” pluggable and chainable to implement the business services
- D6.1 – Testbeds Status and Upgrades - I
  - All the pilots filled a form about the characteristics that will be needed about the testbed infrastructure.

Apart from those main/big contributions, the different pilots are also collaborating with WP3-WP6 (technological WPs) about the different technologies to be used and implemented. Depending on the topic of each task/deliverable, related pilots will contribute or not. Some of this work is already real in the way of deliverables produced by these WPs, meanwhile, other contributions are yet to come. It demonstrates the intensive collaboration and inter-workpackage work in INFINITECH, working everybody as a whole.

This deliverable builds above the base of all these contributions, not (re)defining them again. It rather focuses on reporting the status and readiness of the various pilots.

## 1.3 Structure

For each one of the pilots the deliverable covers:

1. **Preparation and status:** general overview of the pilot and current status. Very quick reference to the user stories, business services, requirements, data sets and testbeds that have already been described in previous deliverables from other WPs (Section 1.2 above)
2. **Technological components and Services:** to point the main building blocks to be developed.

3. **Data sets status:** data sets have been already specified in another deliverable. Just a quick pointer to those previous contributions and focus on the current status of data sets gathering status (size, availability, available through the sandbox, etc).
4. **Testbed:** Status of the testbed if already available.
5. **Others non-technical requirements:** status of other non-technological requirements.
6. **Implementation of a first Proof of Concept:** definition and status of a first Proof of Concept, and how it has been defined in the scope for this first PoC.
7. **Next steps and timeline.**
8. **Conclusions, issues and barriers.**



## 2 Pilot sites preparation

### 2.1 Pilot #1 Invoices Processing Platform for a more Sustainable Banking Industry

The main objective of the pilot is to develop, integrate and deploy a data-intensive system to extract information from notary invoices, in order to:

- Establish the sustainability index of each notary based on the number of physical copies that are issued.
- Provide to financial institutions the information (properly indexed) about the documents that are finally generated by notarial services required by the bank.
- Promote notarial services from those with the higher sustainability score.

The innovation of the pilot lies in the applicability of Artificial Intelligence technologies over scanned physical documents (notary invoices) for cost savings and increased effectiveness. Currently, many physical documents, and copies (some of the redundant), have to be managed. Each physical copy and its control causes significant costs over the period of the financial products lifetime. AI can be leveraged to extract relevant indicators from digitized invoices, which in turn can be used to automatically and accurately rate notaries based on a sustainability index. (More details can be found on D2.13).

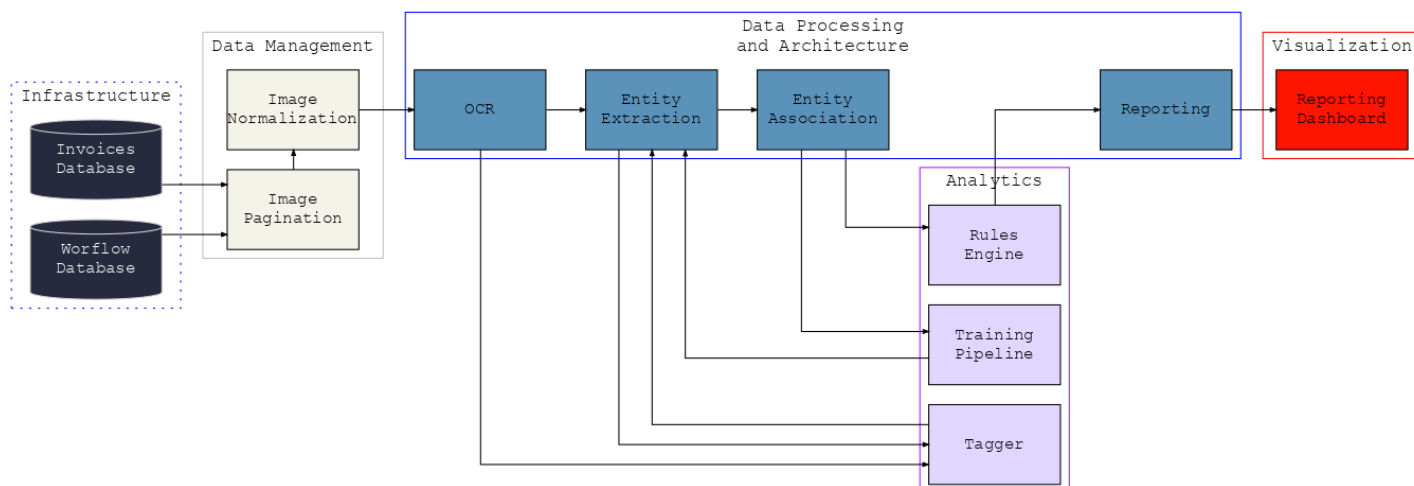


Figure 1: Pilot#1 Reference Architecture

Following a list of partners participating in the pilot and their different roles/contributions:

- Bankia's Auditing department provides the business use case, the functional requirements, and the expert knowledge about entities to extract, business rules, alert generation and information dashboard. It also provides the cloud environment for the deployment of the storage and computation platform. It provides the expert knowledge for invoice tagging and validation of the final product.
- GFT provides the architectural design, platform implementation, algorithm design, training, validation and implementation, together with the document pre-processing. GFT carries out the development of the different components.
- Final users are Bankia's internal auditing department, where a real task will be automatized, thus obtaining a real return of investment and key performance indicators.
- In the pilot will also collaborate: FBK (Fondazione Bruno Kessler) as the data science expert, RB (ReportBrain) as solution expert with expertise in text-analytics and sentimental analysis and INSO (Insomnia Digital Innovation Hub) as business and development advisor.

Due to the modular nature and close interaction between technology and business actors, the solution has been showcased to different European and US potential customers, that have showed a keen interest. Feedback is that, with adaptations with regards to their business workflows and retraining for their documents, both of which is completely feasible, this solution will constitute a compelling technology and business case. Follow-up meetings are scheduled starting from January 2021.

The technology developments have an impact on the implicit training of different actors:

- Business users that have been training in the statistical nature of the results coming from such tools, and therefore the interaction and cooperation human-AI tools.
- Business managers to assess the impact and specificity of the use of such technologies.
- Data scientists and software architects and developers that have been trained during the development of the present project.

### 2.1.1 Technological components and Services

The main technological components that will be implemented and integrated as part of this pilot are:

- Invoices and invoicing workflow database.
- Document ingestion.
- Document pre-processing: document pagination, PDF to image conversion, image normalization, OCR.
- Document entities and region-of-interest extraction: machine learning models and Natural Language Processing extractors for the identification and extraction of entities of interest: billable concepts, prices, headers, addresses, etc.
- Entity association: graph deep neural networks for the identification of related concepts: *e.g.* that a certain billable concept corresponds with an identified price and identified.
- Business rules engine: application of compliance business rules for the generation of alerts and reports
- Data Tagger: for the tagging of training invoices examples
- Document validator: for the verification of processed invoices
- Training and inference orchestrated pipelines.
- MLOps tools: Models and data Repository, code repository.
- Reporting business dashboards and operational databases

### 2.1.2 Data sets status

Data will be extracted from 32.300 real invoices documents from 3.000 different notaries extracted from Bankia systems. Invoice documents will be digitalized in PDF format or may also arrive already digitalized from other channels (email attachments, bulk sftp, etc.). Data type will be: PDF/Image/Text. Data format will be: PDF/ PNG/ TXT.

At the present time, the whole dataset is available in the testbed. The annotation tools, together with the OCR, and a preliminary version of the training and inferencing pipeline are deployed in the testbed (more details in the PoC Sec. 2.1.5).

INFINITE CH Pilot# Dataset Provider	Dataset Name	Dataset (short) description	Owner	License/ Privacy	Anonymized	Capability of Synthetic Data Production	Data Type	Data format	Data store
Pilot#1 BANKIA	Notary invoices	* 32.300 invoices documents, from 3.000 different Notaries	BANKIA	Confidential data (Notary invoices)	No	No. Certain capability of data augmentation introducing rotations, noise, etc..	PDF/Image/Text	PDF/PNG/TXT	Apache Hadoop HDFS, Elastic Search

Table 6 Pilot #1 Data Sets Status

### 2.1.3 Testbed

A cloud-based testbed will be implemented using AWS Bankia Private Cloud, with an estimated volume of data of 2TB (see Table 6). The test bed is already available, and the hardware to be used will depend on the task to be accomplished:

- For training and tagging AWS EC2 instance of the type g4dn.xlarge with 200 GB of disk with GPU.
- For inference, normal computing optimized instances c6g.2xlarge or the same type g4dn.xlarge, with the AWS Deep Learning AMI (Ubuntu 18.04).

Some more details about different tools and software components to be deployed follow:

- Data management: linux file system, S3, elastic search.
- Data processing: kafka, Kubeflow.
- Data analytics and AI related tools: tensorflow 1.5, sklearn, pandas, numpy, seaborn.
- Data tagging: labelme.
- Data visualization: kibana, Floent, Prometheus.

### 2.1.4 Others non-technical requirements

Data privacy for the invoices must be guaranteed during all the development phase.

### 2.1.5 Implementation of a first Proof of Concept

This first Proof of Concept is addressed to the implementation of a two-sided machine-learning based system at scale.

From one side, to automatize the capture and extraction of the unstructured information in scanned documents using computer vision and machine learning deep neural networks. This implies to develop, integrate and deploy a data-intensive system to extract information from notary invoices to establish a sustainability index of notary services based on the number of physical copies issued, that will be used by the bank.

From the other side, to capture the business rules expressed as concept associations (e.g. invoiceable concept + related quantity + related price) that in an unstructured way are scattered along the document with high variability. Finally, automatizing the whole process at scale coupling with automated workflows for document capture and reporting in a real financial institution environment.

Figure 2 shows interactions and workflow, from high level point of view, between the main components. Invoices are automatically ingested by the system to start the processing, yielding the OCRed document, together with the extracted fields, the association between the corresponding fields, and the application of the rules. Later, results will be stored in the ElasticSearch database and finally, the summary accessible by a dashboard.

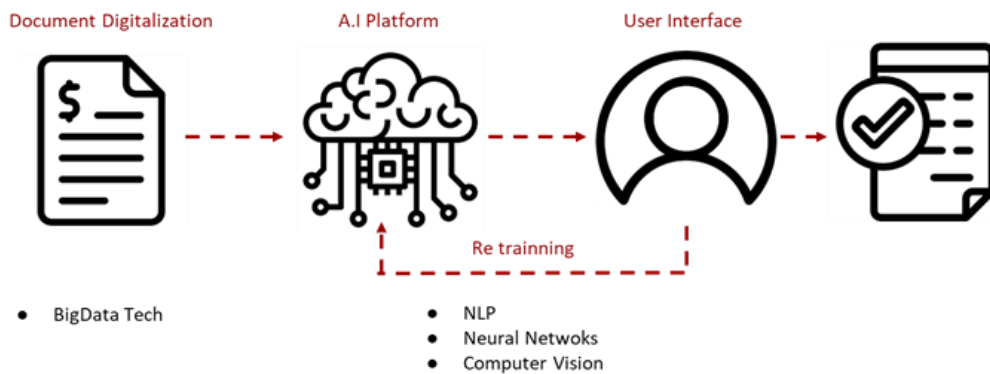


Figure 2: Pilot #1 Main Components Interactions

Data scientist together with domain experts will retrain the system by checking results, in a supervised way.

### 2.1.6 Next steps and timeline

Time line and roadmap is proceeding as planned and is depicted in Figure 3. The plan is divided in two weeks sprints following agile methodology. The project is divided in three different lanes, corresponding to three different technological aspects: the development of machine learning models, the deployment in the corresponding infrastructure and the development of the reporting and business user interface. First milestone is the workshops with the business user to capture the requirements and the study and tagging of the invoices. Also the definition of the reporting dashboards. A second milestone is the development of a first version of the extraction algorithms and their deployment in production. A third milestone is the deployment of the final pipelines. In parallel the reporting dashboards implementation, together with the document storage.

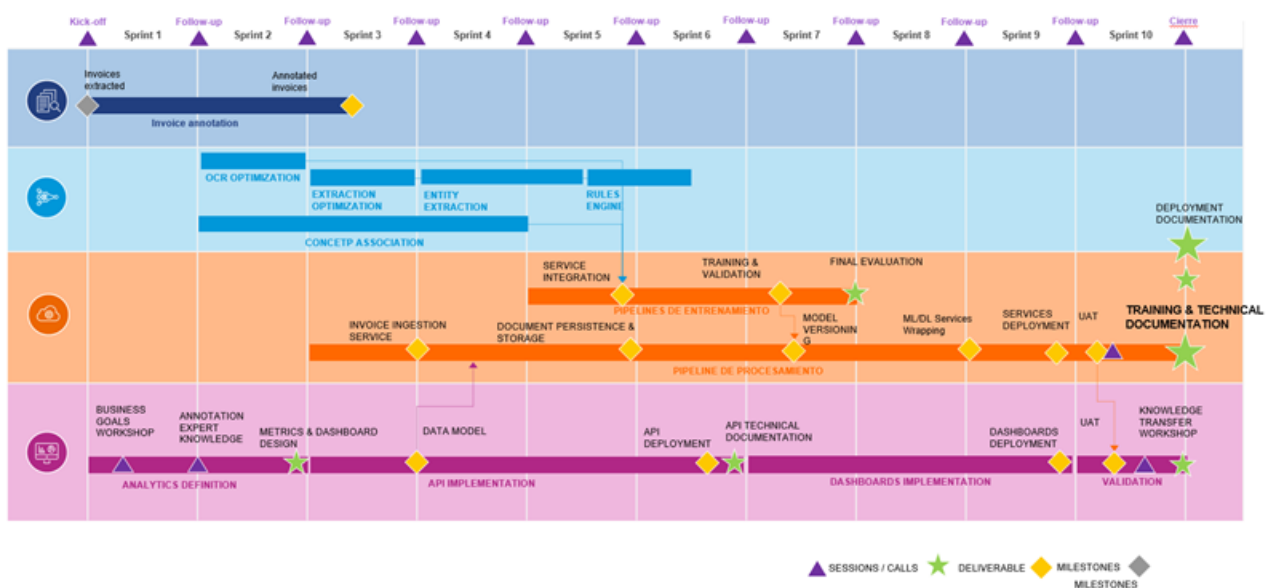


Figure 3: Pilot #1 Timeline

Once deployed, the PoC will be put in use by the auditing department of the bank. The goal is to have a complete automated system able to have a 10 – 20x fold increase in the coverage of the invoices, screening the ones that are non complaint according to the internal procedures.

## 2.1.7 Conclusions - Issues and Barriers

The project is proceeding as planned, with the dataset for training received, tagged and with an initial end-to-end model developed with an initial deployment in the testbed. The testbed is fully operational.

End-users from the auditing department have been involved in the development following an agile methodology. Their roles has been crucial in the:

- 1) Validation of the information to be extracted and the definition of the ground truth for the document samples.
- 2) Elicitation of the operation workflow and reporting dashboards.
- 3) Definition of the business rules for the concepts' association.

Mismatch of end-user expectations and requirements with the actual project implementation has been addressed by the active involvement of the users in the bi-weekly review meetings.

At the present time, the critical path consists in the coordination and adjustment of the different elements of the pipeline, a characteristic typical for projects with intensive use of machine learning algorithms that result in the combination of many moving parts.

The main barriers like the availability of data, tagged data and expert knowledge for problem definition are mainly removed at the present time.

## 2.2 Pilot #2 Real-time risk assessment in Investment Banking

The aim of this use case is to give traders in investment banking a precise and timely indication of the risk of a given portfolio and specifically changes in risk due to market changes or changes of the portfolio. The need of such knowledge comes from operational as well as supervisory requirements that every regulated financial institute must comply with.

Risk assessment is based on a common risk metric - Value at Risk (VaR) - to be calculated and updated in real-time on both, portfolio level as well as for each individual asset. A second risk measure - Expected Shortfall (ES) - indicating not the maximum amount of a potential loss, but the expected loss with a given probability, will be derived at a later stage. The pilot will furthermore implement the evaluation of what-if-scenarios allowing pre-trade analysis, i.e. estimating changes in risk measures before a new trading position is entered. In addition, the pilot will implement a sentiment-based decision support indicator derived from financial and economic news data and social media channels.

The pilot will support institutional traders, asset managers, risk managers and wealth management experts in:

- **Calculating the Value-at-Risk (VaR) of their Portfolios.** Emphasis will be paid on FOREX (FX) portfolios<sup>3</sup>, yet the system will be applicable for other types of portfolios as well.

---

<sup>3</sup> Don Bredin. & H. Stuart (2004). FOREX Risk: Measurement and Evaluation Using Value- at-Risk, Journal of Business Finance & Accounting, 31: 1389–1417.

- **Evaluating what-if scenarios for alternative Portfolios based on their VaR.** In practice the system will simulate alternative investment strategies and will provide relevant information to the end-users to allow them to shape their investment decisions.

The main innovations of the pilot lie in:

- The calculation of VaR at very short timescales based on the processing of high-ingestion data.
- The employment of ML-based VaR calculation techniques that will yield more accurate values and will facilitate traders in better understanding and framing the risks of their portfolios.

## 2.2.1 Technological components and Services

The components to be implemented are depicted in Figure 4: Pilot #2 Data Science Pipeline, which illustrates the data science pipeline for the pilot. They include:

- **Data ingestion component.** Ensures the ingestion of real-time data in the database of the pilot (XLS database). It is destined to cope with the high ingestion rates of the real time data.
- **Market Sentiment component.** Extract market sentiment for specific assets of the portfolio and provides this information to the data to reinforce the accuracy of the VaR calculation and/or to provide alternative methods for VaR calculation. The component is not implemented in the early Proof of Concept that is described in this deliverable.
- **VaR calculation component(s).** Scientific Computing and Machine Learning components, which calculate the VaR of the portfolio based on different methods (e.g., historic method, variance-covariance, monte carlo simulation). They harness data from the LXS dataset.
- **End-User Dashboard component.** Provides user friendly visualization of the VaR parameters for different portfolios owned by the user.
- **Semantic Interoperability component.** Provides an interface for access to FIBO data, while supporting their parsing. The semantic annotation and the structuring of the data according to FIBO is performed in WP4 and hence the relevant description is beyond the scope of this deliverable.

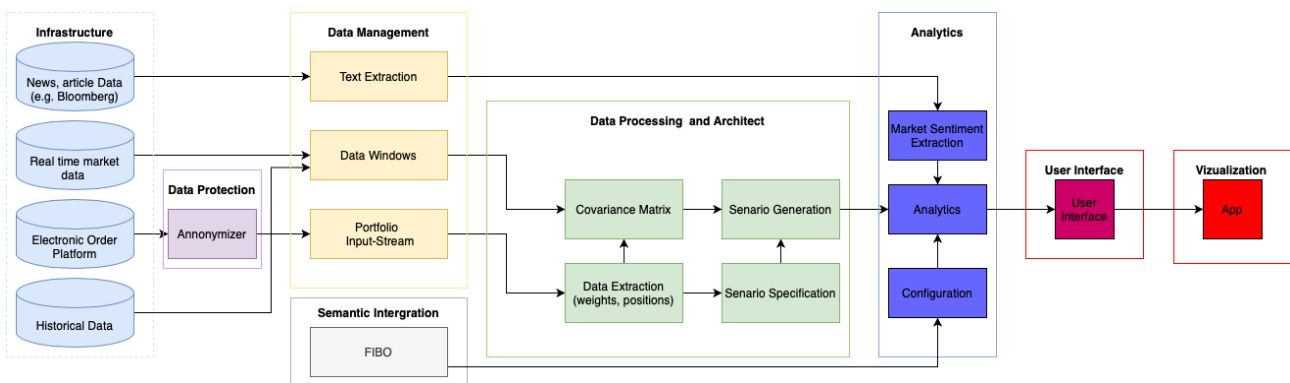


Figure 4: Pilot #2 Data Science Pipeline

The pilot pipeline conforms to the INFINITECH-RA specification i.e. the various blocks are structured according to the modules and layers of the INFINITECH-RA. Likewise, the deployment of the components adheres to the guidelines of the INFINITECH reference testbed.

The implemented pilot's deployment diagram for the first proof of concept can be seen in Figure 5. More specifically:

- The main elements of this deployment are:
  - LXS Database (Docker container) containing Historical Ticker data.

- predict\_var (Dockerized python scripts) for time series pre-processing and VaR prediction
- visualize\_var (Dockerized Flask Web application) to visualize FOREX assets’ historical statistics, VaR predictions and perform What if Analysis.
- New ticker data (test set) are injected from a csv file to the predict\_var docker using kafka in between. In the next version LXS DB will be used instead of a csv file.
- predict\_var docker reads once historical data from LXS DB to be used as a training set for VaR calculation. As new ticker data is created (from the test set) the training set is updated in predict\_var docker.
- The predicted results are written back to the LXS DB.
- visualize\_var read predictions from LXS DB to update dashboards dynamically.

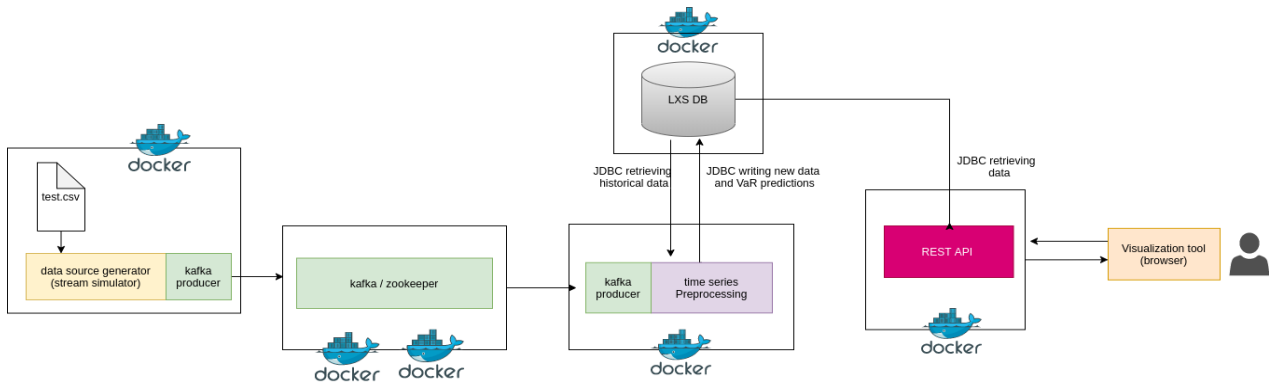


Figure 5: Pilot #2 Deployment Diagram

## 2.2.2 Data sets status

The datasets to be used in the pilot, along with the data collection process is described in detail in deliverable D5.13. It is also summarized below for completeness reasons.

The pilot will leverage FOREX (FX) data provided by the JRC Platform and other trading platforms via Forex APIs. The data will include:

- Trade Data (i.e. data with the assets) of the user that will be used to calculate the portfolio(s) of the user and their Value at Risk.
- Tick Data (i.e. Historical market data) that will be used in the different methods for VaR calculations such as Monte Carlo simulations.
- Alternative data (e.g., data from news) that will be used for market sentiment analysis based on NLP (Natural Language Processing Techniques). News articles and Twitter posts.

Trade Data and Tick Data will contain information such as: the name of the instrument in FOREX trading (ex. GBPUSD for the exchange of GBP to USD), Timestamp that denotes when the trading took place, the Quantity and the Closing Price.

INFINITECH Pilot# Dataset Provider	Dataset Name	Dataset (short) description	Owner	License/ Privacy	Anonymized	Capability of Synthetic Data Production	Data Type	Data format	Data store
------------------------------------	--------------	-----------------------------	-------	------------------	------------	---	-----------	-------------	------------

D7.1 – Report on Pilot Sites Preparation - I

<b>TradesData JRC</b>	Trades data	history of executed orders, making up the current portfolio and updated in real time in order to reflect changes	JRC	proprietary data from JRC model portfolio	No	Yes	numeric	CSV	MySQL
<b>TickData JRC</b>	Tick Data	Existing historical price tick data (every price update on seconds basis) for the most liquid Forex, Indices and Derivatives	JRC	proprietary data from provider	No	Yes	numeric	CSV	MySQL
<b>JRC, INNOV On Line Trading platforms</b>	Derived analysis data	risk measures, correlation matrices	JRC	open	No	Yes	numeric	CSV	
<b>Open Source Data</b>	News articles and Twitter data	Sample Data for Market Sentiment Analysis	JRC	open	No	No	Text	TXT	

Table 7 Pilot #2 Data Sets Details

### 2.2.3 Testbed

The pilot will be deployed at the NOVA testbed. The initial Proof of Concept is deployed in the INFINITECH reference testbed and will be migrated to the NOVA testbed once it becomes available.



## 2.2.4 Others non-technical requirements

At later stages of the pilot experiments with more data will be carried out, based on access to data from other trading platforms (e.g., Forex platforms that provide APIs for different assets). Likewise, the estimations for the open source datasets to be used are subject to revision.

## 2.2.5 Implementation of a first Proof of Concept

The Proof-of-Concept implementation comprises the following components in-line with the pilot data science pipeline:

- Data Ingestion Component implemented within LXS database.
- Scientific computing components in Python that calculate the Value-at-Risk (VaR) using three different methods<sup>4</sup>, namely:
  - **The Historical Method:** This is probably the simplest VaR calculation method. It relies on significant volumes of historical market data (e.g., typically one trading year data for conventional assets and much more than that for hedge funds) to calculate the price changes for all the assets of the portfolio. Accordingly, it calculates the value of the portfolio for each one of the price changes i.e. the value of the portfolio is simulated many times in-line with the number of price changes in the historic data (e.g., approx. 250-260 times for one trading year). These simulated/estimated values for the portfolio can be sorted and used to form a distribution. Then the VaR at a given confidence level (e.g., 99%) is computed as the mean of the simulated values minus the lowest values (e.g., 1% lowest value for the 99% case) in the series of simulated portfolio values.
  - **The Variance-Covariance Method:** This is also called parametric method. It assumes that returns follow a normal distribution, which is a simplistic yet acceptable assumption during normal market conditions. Given this assumption two parameters can be computed i.e. an expected return and a standard deviation for the portfolio. In case of a portfolio with many assets, the standard deviation should consider the correlation in the price changes of the different assets. The latter requires the computation of the covariance matrix of the various assets (i.e. the correlation coefficient of the assets). Based on the mean and the variance of the portfolio its value distribution is calculated and the value at the 95% or 99% confidence interval is produced. The method works quite well when there is a large sample size for the assets of the portfolio, as well as when the distributions of the asset prices are known.
  - **The Monte Carlo Method:** This method develops randomly scenarios for the future price of the portfolio based on some non-linear pricing models. Accordingly, it creates the distribution of these future prices and takes their losses at the target confidence interval. The method is more reliable when dealing with complex portfolios and complicated risk factors. Its advantage compared to the first two methods is that it is not restricted to scenarios seen in the past, but may also consider scenarios more extreme than those contained in the historical data due to its random component and thus is expected to be more realistic.
- The visualization dashboard, which displays VaR Charts for each one of the three methods and two confidence intervals (95%, 99%)<sup>5</sup>. A snapshot of the charts of the dashboard is depicted in Figure 6.

<sup>4</sup> Pilar Abad, Sonia Benito, Carmen López, A comprehensive review of Value at Risk methodologies, The Spanish Review of Financial Economics, Volume 12, Issue 1, 2014, Pages 15-32, ISSN 2173-1268, <https://doi.org/10.1016/j.srfe.2013.06.001>.

<sup>5</sup> Confidence is an important parameter of VaR calculation. For example, if for a given portfolio the 95% confidence of a one-day VAR is 100.000 EUR, this means that there is a 95% confidence that the portfolio will not lose more than 100.000 EUR within a day. 95% and 99% are the most common confidence intervals used in VaR calculation.

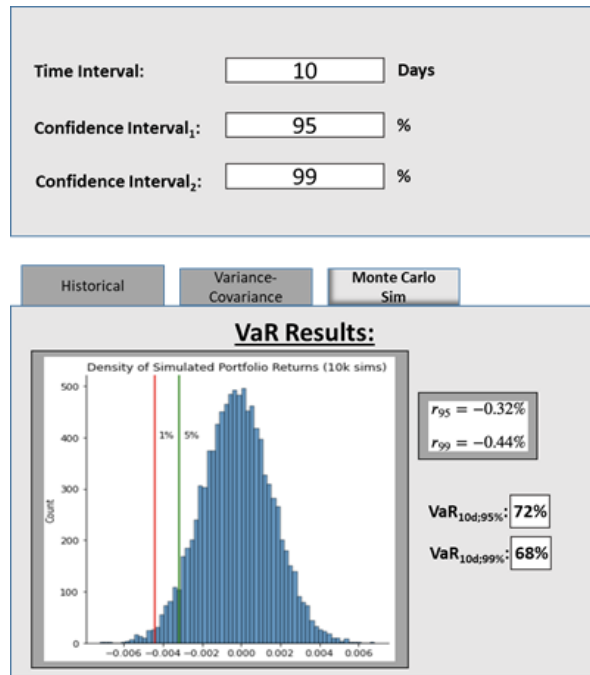


Figure 6: Pilot #2 Dashboard for parameter configuration and visualization

- A comparative visualization view of all three methods with different parametrization and their development over time on a daily basis as depicted in Figure 7.



Figure 7: Pilot #2 Charts for Three VaR Calculation Method

The Proof of concept leverages reduced versions of the “Trade Data” and “Tick Data” datasets i.e.:

- The Trades comprise 3 popular FX assets. The scale-up of the Proof-of-Concept will support more complex portfolios.
- The Tick Data comprises historic data about the corresponding Forex assets in the period March 2020-October 2020. It is considered a sufficient dataset for the Proof-of-Concept and the validation of the various methods. However, the scale-up of the pilot will use more data and will experiment with different historic windows.

## 2.2.6 Next steps and timeline

The next steps in the development of the pilot include:

- **Development of first integrated prototype with Configurable Portfolios (M13-M15):** This will evolve the proof of concept implementation to operate based on multiple dynamic and configurable portfolios, rather than based on static portfolios. An initial implementation of the What-if Analysis will be also provided. Likewise, the end-user dashboard will be updated in-line with the latest requirements from JRC.
- **Migration of the Integrated Prototype in the NOVA Testbed (M14-M16)<sup>6</sup>:** As part of this step the pilot prototype will be migrated to the NOVA testbed/sandbox, from the INFINITECH testbed. The step assumes that the testbed of NOVA will be available.
- **Fine-Tuning of the prototype and Stress Testing (M16-M18):** As part of this step, the pilot MVP will be fine-tuned based on feedback from JRC and other relevant end-users/stakeholders. Improvements in the dashboard and validation of the VaR calculation techniques are foreseen. Moreover, the system will be scaled up and tested with larger volumes of data. Relevant progress will be documented in Deliverable D7.2.
- **VaR Revisions -Integration of Market Sentiment and Semantic Interoperability Modules (M18-M27):** In this step the market sentiment module will be integrated, while more and different VaR calculation methods will be supported. Likewise, the pilot will support integration with the semantic interoperability module, including support for the FIBO semantic format. At the end of this phase a fully-fledged version of the pilot system will be available at TRL=5-6. It will be included/reported in D7.3.
- **Technical and Business Validation (M18-M27):** In the scope of this phase the pilot will be fine-tuned based on feedback from stakeholders. Both technical (e.g., performance metrics, accuracy of VaR calculation) and business metrics (e.g., system cost, ROI/IRR calculation) will be calculated and assessed. Based on feedback from the validation process the final (TRL=7) system will be developed and described in deliverable D7.4.

## 2.2.7 Conclusions - Issues and Barriers

The implementation progresses smoothly in terms of its BigData and data analytics parts. Nevertheless, the datasets used are still quite limited. Furthermore, the prototype of the market sentiment component is not available. Likewise, the NOVA testbed is not fully operational. These are two of the main risks that have to be monitored and cleared in the coming months i.e. within the period M13-M18, so as to ensure that the full-scale implementation is on track.

In collaboration with WP4, this pilot is enhanced with semantic interoperability features/functionality, which were not originally foreseen. Specifically, the pilot systems will support inputs (e.g., Trades Data) in FIBO (Financial Industry Business Ontology)<sup>7</sup> semantic format, to support VaR calculation in cases of portfolios for large investors (e.g., large investment banks, institutional investors) that might hold assets/trades across multiple platforms. In this case, the VaR of a portfolio might have to be calculated based on data from multiple platforms that produce data in different semantics and formats. FIBO will ensure the semantic integration and semantic interoperability of these streams/trades towards facilitating VaR calculation for large portfolios. The integration of a semantic interoperability module in the pilot system is considered as a highlight for the pilot.

---

<sup>6</sup> At the time of writing the NOVA testbed is estimated to be up and running during M15

<sup>7</sup> <https://spec.edmcouncil.org/fibo/>

## 2.3 Pilot #3 Collaborative Customer-centric Data Analytics for Financial Services

Pilot 3 has been affected from some internal changes in their partners. They are working in re-defining and finalizing the scope of the pilot. More details for this Pilot will be included in the next version of this deliverable.

## 2.4 Pilot #4 Personalised Portfolio Management

The main goal is to develop and adapt, within Privé Managers Wealth Management Platform, an optimization algorithm (further on called Privé Optimizer “AIGO”), and an artificial intelligence engine to aid investment propositions for retail clients.

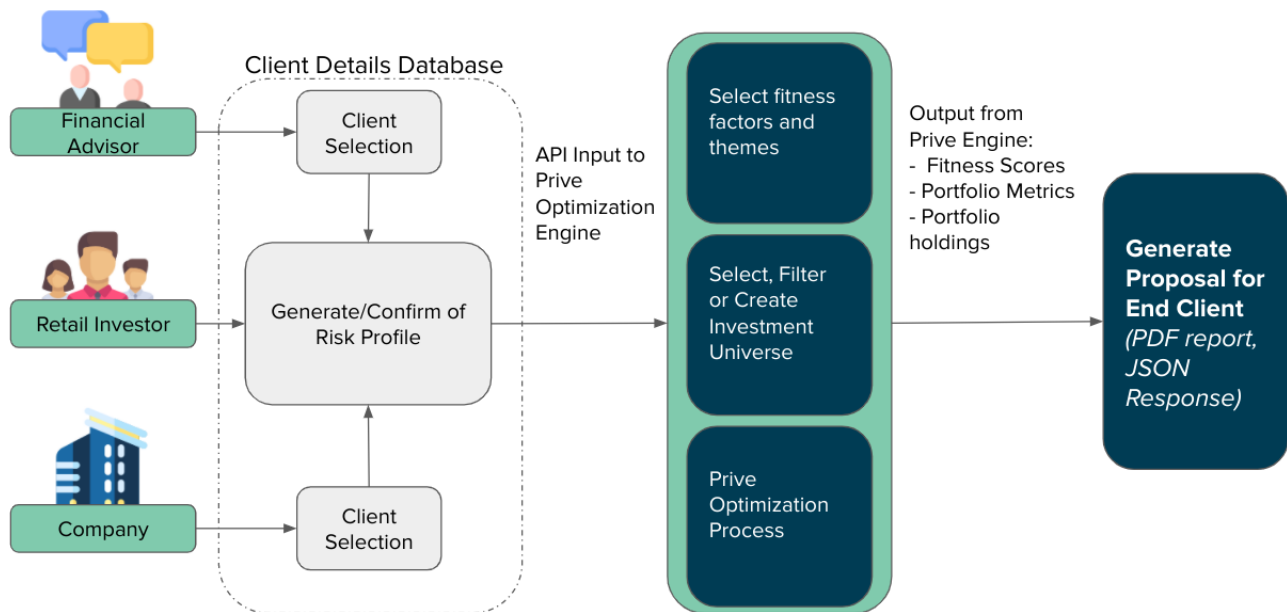


Figure 8: Pilot #4 Roles and Services

This pilot will explore the possibilities of AI Based Portfolio construction for Wealth Management, regardless of the amount to be invested (i.e. “Private Banking could be for everyone”). The AI Based Portfolio Construction will enable advisors and/or end-customers, to use the existing Wealth Management Platform “Prive Managers” and make use of its risk-profiling and investment proposal capabilities, starting from his/her personal risk-awareness. AIGO (or Prive Optimizer) allows for a variety of use cases which cater to the needs of financial advisors, end-clients and financial firms alike.

Starting from a client’s cash pool or current investments/portfolios, the user will select the fitness factors and constraints or preferences to perform the portfolio construction, based on the client’s risk profile and preferences. The optimisation tool that will be developed from the Pilot, will run on a pre-set universe of assets taking into account all the input data and constraints. The AI genetic algorithm will generate a new proposal, where the selected preferences and risk parameters have been recognised. The optimisation tool can be run multiple times, after the necessary changes in initial parameters are made. In this context the main innovation of the pilot lies on the applicability of AI technologies to build customized portfolios (Private Banking for everyone).

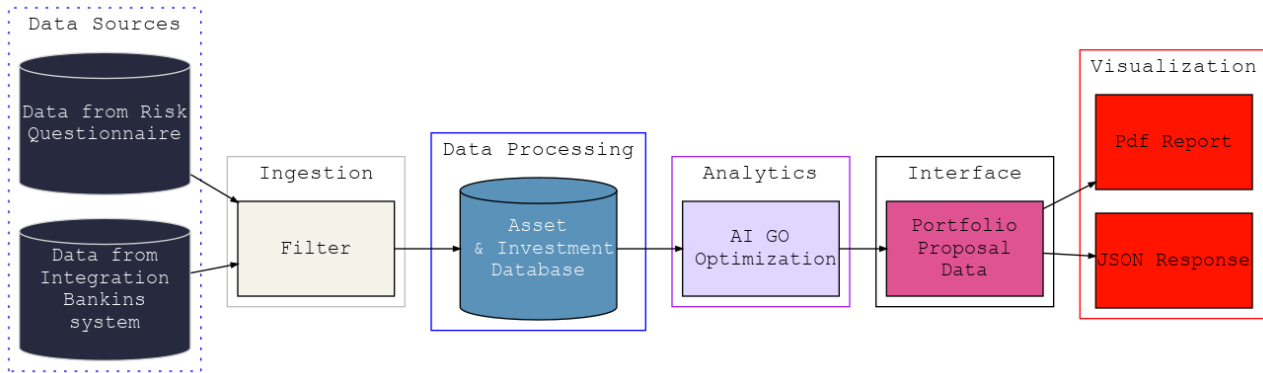


Figure 9: Pilot #4 Reference Architecture (From D2.13)

Pilot’s Reference Architecture (Figure 9) and main data flows have been presented and detailed in D2.13. This RA can be simplified considering:

- A Data Management layer, that performs data ingestion based on cash pool or current investments/portfolios data, quality checking and harmonization of the data provided in order to be imported into the datastore for use of the pilot’s functionalities.
- A Data Processing Layer in charge of homogenise and store all data collected, according specific data models to perform the data preparation for portfolio construction, based on the client’s risk profile and preferences.
- An Analytics layer, that will be based on the AIGO optimisation tool that will be developed from the Pilot, which will run on a pre-set universe of assets considering all the input data and constraints. The AI genetic algorithm will generate a new proposal, where the selected preferences and risk parameters, based on the data provided from the customer and the relative investments /portfolios, available.
- Finally, a visualization layer will provide the proposed portfolio suitable for the specific customer through a report in PDF format or JSON response.

Privé external stakeholder regarding AIGO is currently Report Brain. Privé will provide the technology for the optimization process. On top of that Reportbrain will support Privé with their own specific dataset. In that way, the development will be carried out by Privé with the support of Reportbrain. The end user will be advisors, asset managers, insurance companies, banks, family offices or their end-users/clients.

## 2.4.1 Technological components and Services

The High-Level Architecture presented in Figure 9 presents the software components that build the Pilot’s use cases. This figure has been used in D6.1 Testbeds Status and Upgrades to identify hardware requirements, and in D2.5 Specifications of INFINITECH Technologies to describe the technologies behind its principal components.

This document links the shown software components with the corresponding Reference Architecture layers, providing some details about their implementation. In this sense:

- Data Collection (Data Management layer of the RA) based on customers cash pool or current investments/portfolios data
- Customers’ & investments/portfolio Data quality check (Data Processing layer in RA): according to specific data models, in order to perform the data preparation for portfolio construction, based on the client’s risk profile and preferences.
- AI Based Portfolio Optimization Process (AIGO) that will be developed from the Pilot based on AI Algorithm, will run on a pre-set universe of assets taking into account all the input data and constraints, generating a new proposal, where the selected preferences and risk parameters for a specific customer.
- New proposal for the personalized portfolio will be visualized through a PDF report generation or a JSON extract that will be able to be imported in any relevant portfolio management tool.

Both inputs and outputs will be stored in Privé own cloud. AI fitness-functions within the AIGO will be callable via API based on the initial user preferences inputs. All the datasets will also be stored on Privé side for both inputs and outputs for the algorithm.

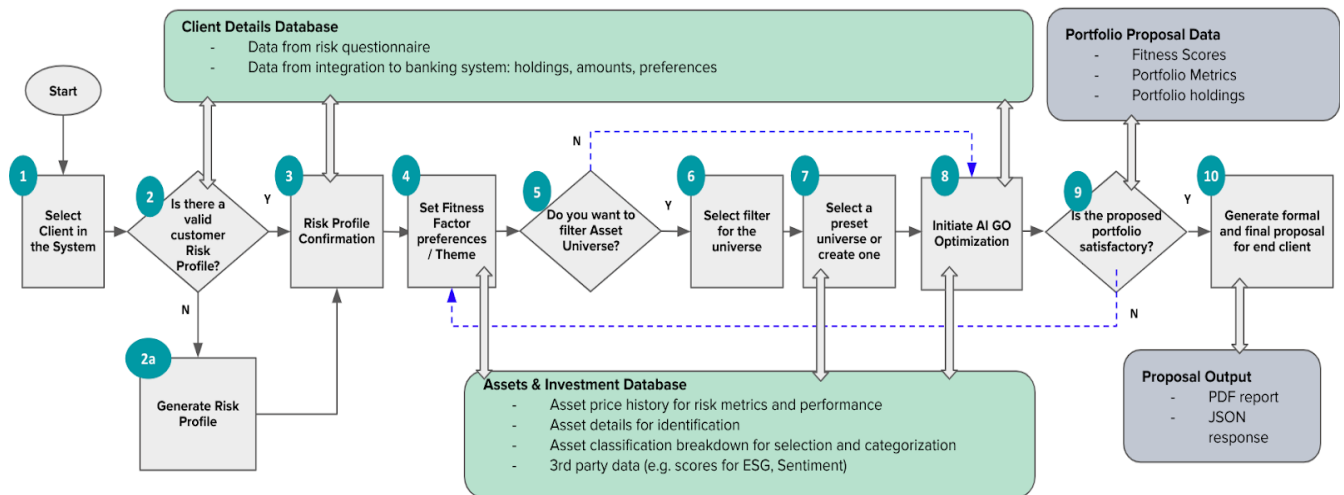


Figure 10: Pilot #4 Process flow diagram

Starting from a client’s cash pool or current investments/portfolios, a risk profile is created or an existing one is updated (Steps 1 to 3 on Figure 10). Then the user will select the fitness factors and constraints or preferences to perform the portfolio construction, based on the client’s risk profile and preferences (Step 4). The optimisation tool that will be developed from the Pilot, will run on a pre-set universe of assets taking into account all the input data and constraints (Steps 5 to 7). The AI genetic algorithm will generate a new proposal, where the selected preferences and risk parameters have been recognised (Step 8 and 10). The optimisation tool can be run multiple times, after the necessary changes in initial parameters are made, based on that the proposed portfolio is satisfactory or not (Step 9).

### 2.4.2 Data sets status

The datasets available for the first proof of concept used by this pilot will be Financial Market Price Data and Financial Market Asset Master Data. Below indicated the whole Data set planned for the entire project for this pilot. The ones marked with “yes” will be used in the PoC.

Dataset Name	Dataset description	Data format	Anonymization	Already available in the first proof of concept
Customer Transactions Data	Customer Transactions Data fetched directly from the Bank or an Asset Manager. It consists of customer securities and cash transactions through their deposit accounts.	CSV	Confidential Data	No

Financial Market Price Data	Financial Market Price Data fetched from several Market Data Providers. It consists of price data for Stocks, Bonds, Mutual Funds and or other assets like certificates/warrants.	Txt	Open, partially license agreements with data providers needed	Yes
Financial Market Asset Master Data	Financial Market Asset Master Data fetched from several Market Data Providers. It consists of asset related characteristics (e.g. expiration date, minimum investment amount, asset class breakdowns).	Txt	Open, partially license agreements with data providers needed	Yes
Customer Risk Profile Data	Customer Risk Profile Data fetched directly from the Bank or an Asset Manager. It consists of customer Risk Profile Data through their account data and profiling, based on B2B customers parameters.	CSV	Confidential Data	No
Mutual Fund, ETF and Structured Products Breakdown	Mutual Fund, ETF and Structured Products Breakdown fetched from several Market Data Providers. It consists of asset breakdowns based on bank data or market data providers breakdown.	CSV	Open/Confidential data, partially license agreements with data providers needed	No
Customer Economic Outlook	Customer Economic Outlook fetched directly from the Bank or an Asset Manager based on questionnaires and Customer Profiles.	CSV	Confidential Data	No

Table 8 Pilot #4 Data Sets Summary

All datasets will be stored within the Privé SaaS solution in a cloud setup. Asset data and Client data are fetched from 3rd party databases and partially from selected market-data providers. The following list describes in more detail the data that will be used from Privé Optimizer or “AIGO” during processing:

- Asset Price History for Risk Metrics and Performance.
- Asset Details for Identification.
- Asset Classification Breakdown for Selection and Categorization.
- Financial Market Price Data fetched from several Market Data Providers.
- Financial Market Asset Master Data fetched from several Market Data Providers.

The output data consists of the single portfolio holdings, their weights and amounts to decide about the Proposed Portfolio. Fitness Factors Scores and Total Fitness Score will be output for both the current and proposed (optimised) portfolio. For both Portfolios also Risk and Return metrics will be shown.

### 2.4.3 Testbed

As indicated below, Privé will be storing its Testbed on its own Amazon Cloud in AWS with an architectural setting as indicated below.

#### Technical Specifications from Privé’s internal Testbed

Hardware Specifications (in case of Cloud Installation, include the relative cloud configuration)

3 VM instances, each with the following:

CPU: Intel Xeon 3 GHz or faster

Core: minimum 2 Core 4 threads

Memory: 32 GB DDR4 1600 or 1866

Hard Disk: 16 GB SSD

#### Technical Architecture Diagram

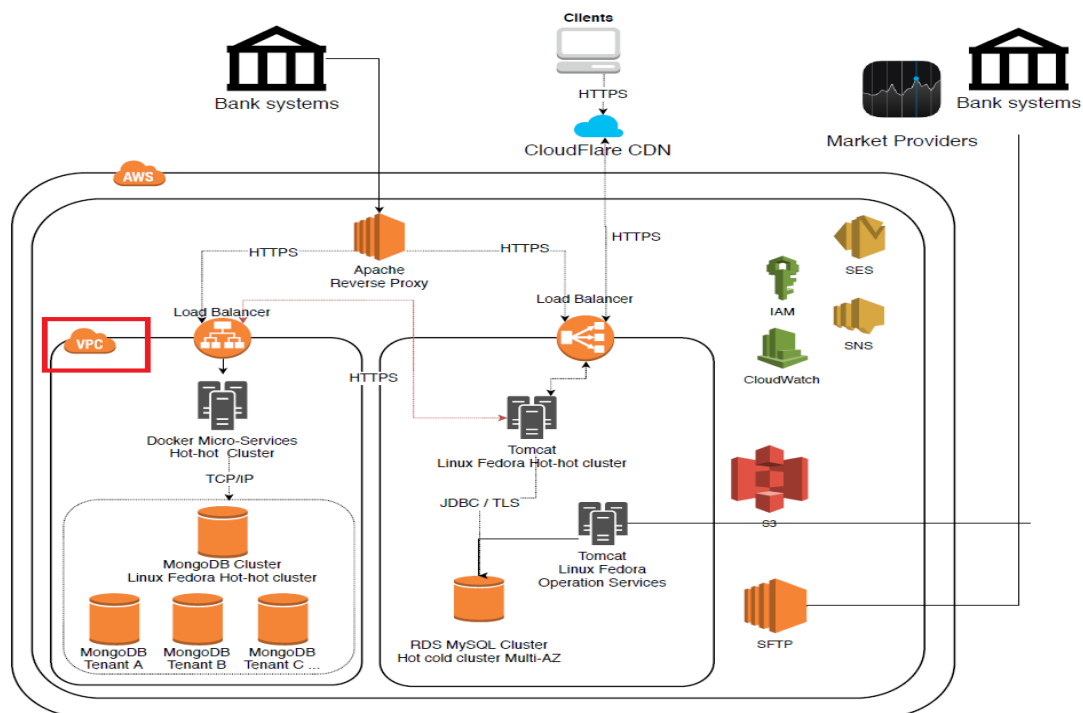


Figure 11: Pilot #4 High-Level Architecture

Software Specifications including for each module the relative software stack that is used (e.g. Operating System, Layered Software, Application Software, Development Platform, Database, etc)

#### Operating System / Application Software / Layered Software

The SaaS platform runs in multiple data centers with active-active setup to achieve high availability. Privé has the following environments: DEV, SIT, UAT and PROD. Data can be transferred via SFTP, FIX or API. Most Privé APIs are REST, but SOAP and GraphQL are also supported. The architecture is based on microservices.



Operating System: Ubuntu 18.04 LTS  
Framework: SpringBoot: 2.2  
Application Server: Tomcat: 7.0.103  
Database: MySQL: 5.6.47  
Database: MongoDB: 3.6  
Language Runtime: Java: OpenJDK 8u242

### Development Platform

We use html5/ReactJS for frontend. Our platform is written in Java, with Spring MVC, Spring boot, and hosted with Apache Tomcat.

Note that The requirements for Hardware specifications (e.g. RAM, No of CPUs, etc) will be required to be defined based on the requirements of the relative technology solutions (e.g. Data Management & Processing, Analytics & AI, etc) that will be used for each Testbed and relative sandboxes, in cooperation with the relevant Technical Partners.

Privé testbed is ready to be used and the tests conducted on our own AWS cloud were successful. The proof of concept will be delivered and presented showing the current API functionalities and results stored in this testbed.

## 2.4.4 Others non-technical requirements

No non-technical requirements will be needed.

## 2.4.5 Implementation of a first Proof of Concept

As a minimum-viable-product or better first Proof of Concept, Privé will be presenting the back-end /calculations capability from the AI GO (Artificial Intelligence Portfolio Construction Optimizer) via API Calls. This will consist of the optimization process presented for an example-portfolio via a couple of so-called fitness-factors which will allow to optimize a pre-given portfolio with a pre-determined investment universe. For the first proof of concept not all services or data sets described in the user stories will be implemented. The figure below highlights the PoC main components:

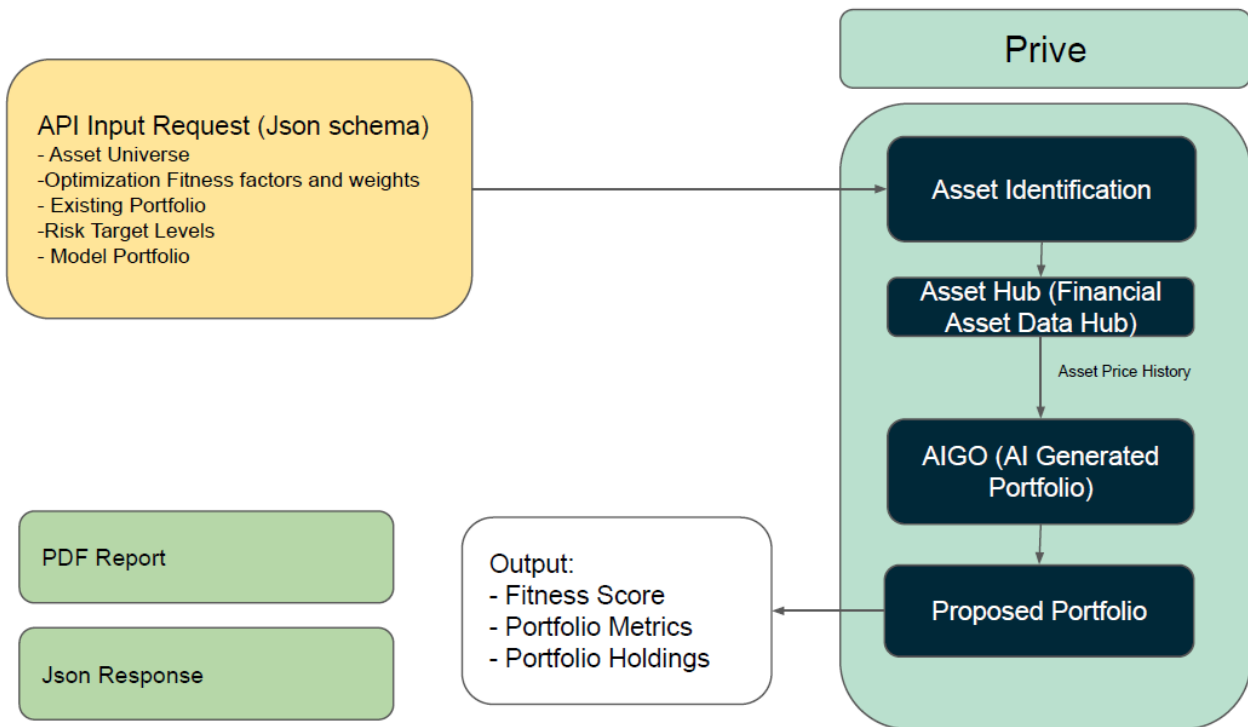


Figure 12: Pilot #4 Components

In this case the input will consist of an investment universe of 50 European stocks and a pre-defined portfolio-example and the out-put will result into an optimized portfolio based on the selected user preferences (fitness-factors functions for the optimizer for that matter).

Both data sets and testbed infrastructure have been described in more detail above. All the inputs and outputs will be callable via API.

### 2.4.6 Next steps and timeline

In the short-term, the next milestones could be set:

- After successfully testing a pair of fitness-factors, the next step will be to add more fitness-factors to the optimization process until the end of the year, Q4 2020.
- Also, potentially another “Market Sentiment” Fitness-Factors based on Report Brain input could be added until the end of the year, Q4 2020
- More functionalities/optimization capabilities will be available for the end of Q1 2021.
- The next big step could be implementing those optimization processes via API in a Front-End.

### 2.4.7 Conclusions - Issues and Barriers

After development started, Privé successfully finished implementing a First Proof of Concept in a UAT Environment stored on our own AWS Cloud. The pilot testbed is already set up and available via SaaS access.

The main challenges consisted of the Market Data Availability Setup on our UAT Environment, as will be required the relative datasets to be enhanced either with more customer portfolio data, or with more variety of financial instruments data available from various sources that will affect the fitness factors and constraints to perform better proposed portfolio construction.

Similar challenges could arise in the future as other investment universes or market providers are made available for the optimization process.

Also, the integration of a so-called new fitness-factor based on Reportsbrain Market Sentiment Factor as an external provider via API could bring up some challenges too, as will require further exploration of AIGO optimisation tool capabilities in order to provide better results for personalized proposed portfolio taking in account also sentiment analysis factor.

In general the outcome of this pilot will be develop and adapt within SaaS based Privé Managers Wealth Management Platform a Portfolio Optimization algorithm AIGO (or Privé Optimizer) , as well as improving and expanding its capabilities as an artificial intelligence engine to support better investment propositions for retail clients, that can be used as SaaS service through an API for other interested parties (investment firms, private banks, wealth management firms, etc).

## 2.5 Pilot #5b Business Financial Management (BFM) tools delivering a Smart Business Advise

Pilot 5b aims to assist Small and Medium sized enterprises (SMEs) clients of Bank of Cyprus in managing their financial health in the areas of cash flow management, continuous spending/cost analysis, budgeting, revenue review and VAT provisioning. In this direction, the Pilot provides a set of AI powered Business Financial Management tools, harnessing available data to generate personalized business insights and recommendations for the SMEs. The main innovation of the pilot lies in the provision of automated and predictive assistance to SMEs about how to plan future liquidity needs and fulfil their obligations.

Main stakeholders of the pilot development include Bank of Cyprus (BOC) and University of Piraeus Research Centre (UPRC). BOC is providing a variety of data mainly regarding its SME clients and their respective transactions, while also being the key driver in designing the Business Financial Management toolkit, which will generate valuable insights and add value to the existing online services for SME beneficiaries. UPRC is working closely with BOC in designing all provided services. It is responsible for the development of all required ML/DL algorithms of the pilot and the technical support of the pilot's implementation throughout the project.

The pilot aggregates a variety of data related to SMEs accounts from Bank of Cyprus' operation data warehouse, which include: (i) account, (ii) customer and (iii) transaction data. Moreover, (iv) open banking data will be utilized to provide a holistic approach, as well as (v) invoice data from the SMEs in order to provide accurate reconciliation services.

The datasets used, as well as the pilot's RA is illustrated in Figure 13. All personal and sensitive data related to SME customers of BOC will be pseudonymized at the bank's premises using a tokenization approach before streamed to the INFINITECH ecosystem to ensure the protection of vital SME data. A reverse pseudonymization will be applied before presenting the data to the SME end user. The RA of the pilot, as included in deliverable 2.13 is depicted below:

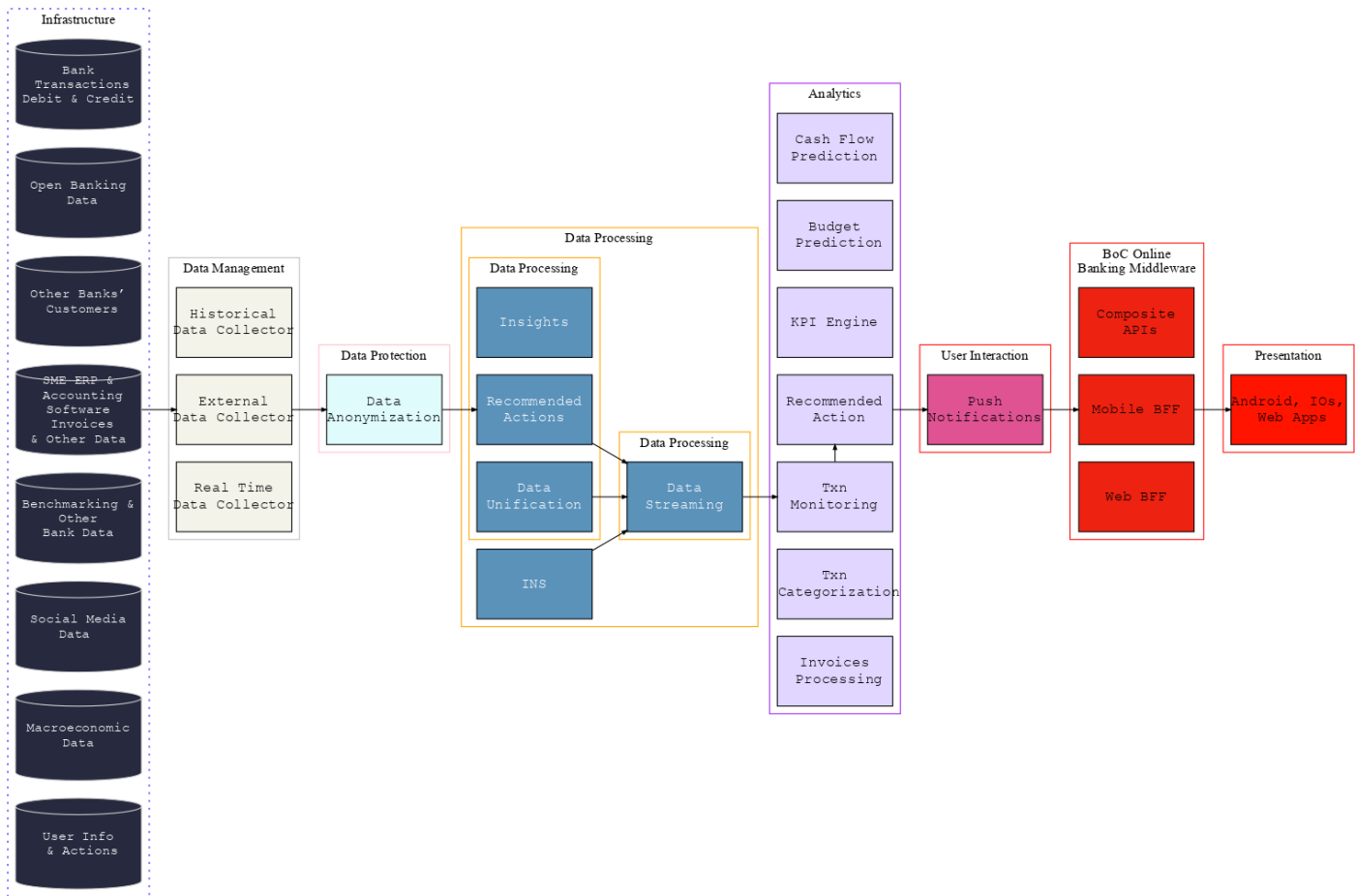


Figure 13: Pilot #5b Reference Architecture

The various components will be containerized using Docker, and a LeanXscale database will be used to store and query the results of the analytics processing, as well as insights generated by the recommender engine. Most of the data analytics components are developed using Python data analytics and ML/DL libraries, i.e. Numpy, Pandas, ScikitLearn and Tensorflow, where data streams required in some components for real time analytics will be handled with Apache Kafka. For the time, a static approach has been followed and all development progress has been done in offline mode in University of Piraeus premises, with all progress being migrated to the INFINITECH ecosystem once the pilot’s AWS testbed is set.

### 2.5.1 Technological components and Services

All services developed focus on providing valuable business insights and recommendations to the SMEs, empowering them to effectively monitor cash flow, budgeting, revenue and perform reconciliation activities, all leading to improved business management and data-driven decision making. The services provided are depicted in the figure below:



Figure 14: Pilot #5 Services

The figures show a set of different services/components/engines. Each one, in a different development stage.

An early version of the **Transaction Categorization Engine**, which is considered a key component, has been developed. This component is in charge of labelling the transactions of selected SME customers of Bank of Cyprus into 20 main categories (with around 80 respective subcategories to be implemented soon). This first version has been implemented combining rule-based classification and ML algorithms.

The development of the **Cash Flow Prediction component** has also been initiated, exploring a variety of ML models to predict the expenses of certain categories of a given account in a short period of time.

These two have already started the development and will be included into the first PoC. The development process will include/add new components:

- Budget Prediction engine that allows setting easily budget targets through the provision of suggested target values as well as simple budget monitoring.
- KPI engine leading to valuable insights on the SME financial health and performance.
- Transaction monitoring engine that watches out for potential anomalies and savings.
- Invoice Processing engine that generates meaningful invoice background info to other components (e.g. Cash Flow Prediction) and SMEs. This applies if respective data can be obtained from SME relative ERP system.
- Benchmark engine supporting comparisons to other SMEs with similar profiles and
- Recommender engine generating actionable insights for a SME that will allow to perform better.

## 2.5.2 Data sets status

All data originating from Bank of Cyprus have already been aggregated internally in the Bank and are already available to the technical partners, mainly UPRC which is developing the two initial components for the PoC, namely Transaction Categorization and Cash Flow Prediction components.

Transaction, customer, and account data related to the respective SME will be drawn from BOC's repository by a real time/historical data collector. Transaction and account data from Open Banking (PSD2), as well as BOC customer data, will utilize a historical data collector. In addition, an external data collector will also be used in order to integrate other related Open Banking/macroeconomic data. At this point, without utilising a functional testbed, a static approach has been used to process available data and develop the first pilot's components.

**Already available datasets:**

- Transaction Data from BOC: a .csv file with around 500MB and 3.5 millions of transactions between 2018 and 2019.
- Accounts Data from BOC: maps accounts with the transactions.
- Customer Data from BOC: links customer to accounts and the available NACE code is used in the transactions' categorization model.

**Datasets available in the near future:**

- Transaction Data from Open Banking (i.e. PSD2 data).
- Other Data (Market).
- Transaction Data from SMEs (optional).
- Other Data from SMEs (optional).

### 2.5.3 Testbed

Bank of Cyprus (BOC) is developing an AWS testbed, based on the technical requirements and guidelines of the relevant partners, and tailored for the unique pilot's components and the required data ingestion. As the testbed's specifications have not yet been finalised and certain bank processes require time, until the bank's AWS ecosystem is available the pilot's first components will be hosted in GFT's AWS environment.

### 2.5.4 Other non-technical requirements

The pilot's component providing competitive advantage among other available BFM tools is considered to be the Smart Virtual Advisor that leverages extensively supervised and unsupervised machine learning, takes into consideration the output from all BFM tools to come up with a holistic view of the SME business corresponding accurate business advise and reconciliation all fostering an optimal day to day business operation. The main non-technical requirement to achieve this will be solving all consent and data protection issues arising from including such enterprise data.

### 2.5.5 Implementation of a first Proof of Concept

The Proof of Concept is aiming to establish the foundation for the various smart Business Financial Management (BFM) engines. To achieve this, the design, development and implementation of a Transaction Categorization engine is prioritized as it maintains a vital role for the development and interconnection of all other components. To demonstrate the integration between the various engines, a basic Cash Flow engine will also be implemented.

The pilot's testbed will be accommodated by Bank of Cyprus, which is going to provide an AWS environment for the various pilot's components and operation. As the testbed development has not yet been completed, the PoC version represents a static development approach, where data have been collected and pre-processed by BOC and then sent to UPRC, where the Transaction Categorization and the Cash Flow prediction components are developed locally at the university's premises in an offline environment. Once development of the testbed is completed, the two components will be migrated to the INFINITECH ecosystem and will be fine-tuned accordingly.

Pseudonymized data have already been transferred to UPRC in .csv format to initiate the development of the two main components of the PoC version. Those data include:

- Customer Data from BOC: Data regarding selected SME BOC clients that will be the first
- Account Data from BOC: Information regarding more than a thousand accounts linked to the abovementioned selected SME clients
- Transaction Data from BOC: Dataset with approximately the transactions of the selected SME BOC clients over the last three years. The dataset is considered the main source for developing the first two pilot's components

Rest of the pilot's data included in 2.5.2 will be utilised to enrich and refine the Categorization and Cash Flow Prediction components included in the PoC and will also be crucial for the development of the rest of the components.

### 2.5.6 Next steps and timeline

Once the PoC has been presented, next steps include:

- Developing BOC's testbed and migrating completed components (Feb 21').
- Developing and integrating the required data streaming/data collecting components (Feb 21').
- Refining and enriching existing components, namely Transaction Categorization and Cash Flow Prediction engines (Jan 21').
- Finalising and developing rest of the Pilot's components for an MVP (Jul/Aug 21').

### 2.5.7 Conclusions - Issues and Barriers

Concluding the Pilot's development is progressing based on the project's timeline already establishing the Transaction Categorization and Cash Flow Prediction components that are considered the foundation for designing and developing the rest of the AI powered components included in the BFM toolkit that will be the outcome for SMEs. Next pilot's milestone is moving all development progress to the cloud environment and setting the required data streaming/data collection mechanisms. Main challenge is the AI powered Business Financial Management tools development and their efficiency, as will be based on the availability of all the required data for SMEs from BOC or from the SMEs in order the final goal to be achieved.

Main goal of the Pilot aims to assist SME clients of Bank of Cyprus (BOC) in managing their financial health in the areas of cash flow management, continuous spending/cost analysis, budgeting, revenue review and VAT provisioning, all by providing a set of AI powered Business Financial Management tools and harnessing available data to generate personalized business insights and recommendations. Machine learning

algorithms, predictive analytics and AI-based interfaces will be utilized to develop a kind of smart virtual advisor with the aim to minimize SME business admin effort, to focus on growth opportunities and to optimize cash flows performance.

## 2.6 Pilot #6 Personalized Closed-Loop Investment Portfolio Management for Retail Customers

Pilot #6 focuses on providing personalized investment recommendations for the retail customers of the bank. National Bank of Greece (NBG) will leverage large customer datasets and large volumes of customer-related alternative data sources (e.g., social media, news feeds, on-line information) in order to make the process of providing investment recommendations to retail customer **more targeted, automated, effective, as well as context-aware (i.e. tailored to state of the market)**. The latter is the main innovation of the pilot. An overview of Pilot #6 is given in the following figure:



Figure 15: Pilot #6 Personalized Closed-Loop Investment Portfolio Management for Retail Customers

Pilot’s Reference Architecture (Figure 16) and main data flows have been presented and detailed in D2.13. This RA can be outlined in three main layers to be implemented through different software components. These main three layers are:

- A Data Management layer, that performs data quality checking and harmonization of the data provided from NBG and imported them into the datastore for use of the pilot’s functionalities. On this first stage, the data that will be used transactions data of deposit accounts, cards, investments and CRM data for a small subset of NBG Customer will be used,
- A Data Protection and Data Processing Layers in charge of cleanse, homogenise and store all data collected, according specific data models provided from NBG operational DWH, so these are available for the analytics processes. Here are also included all the operations needed to anonymise (if required) the captured data and protect this information from unauthorised accesses.
- An Analytics block, fed by the data layers, where different ML/DL technologies and visualization tools will enable data monitoring, analysis and exploitation. Two main AI models will be developed here, the Customer Risk Profile Clustering and Personalized Investment recommendation decision support, that will utilize also the Sentiment Analysis data provided from RB relative engine, in order to provide



for a customer, the recommended products to invest through a visualization application (in the RA’s Visualization layer).

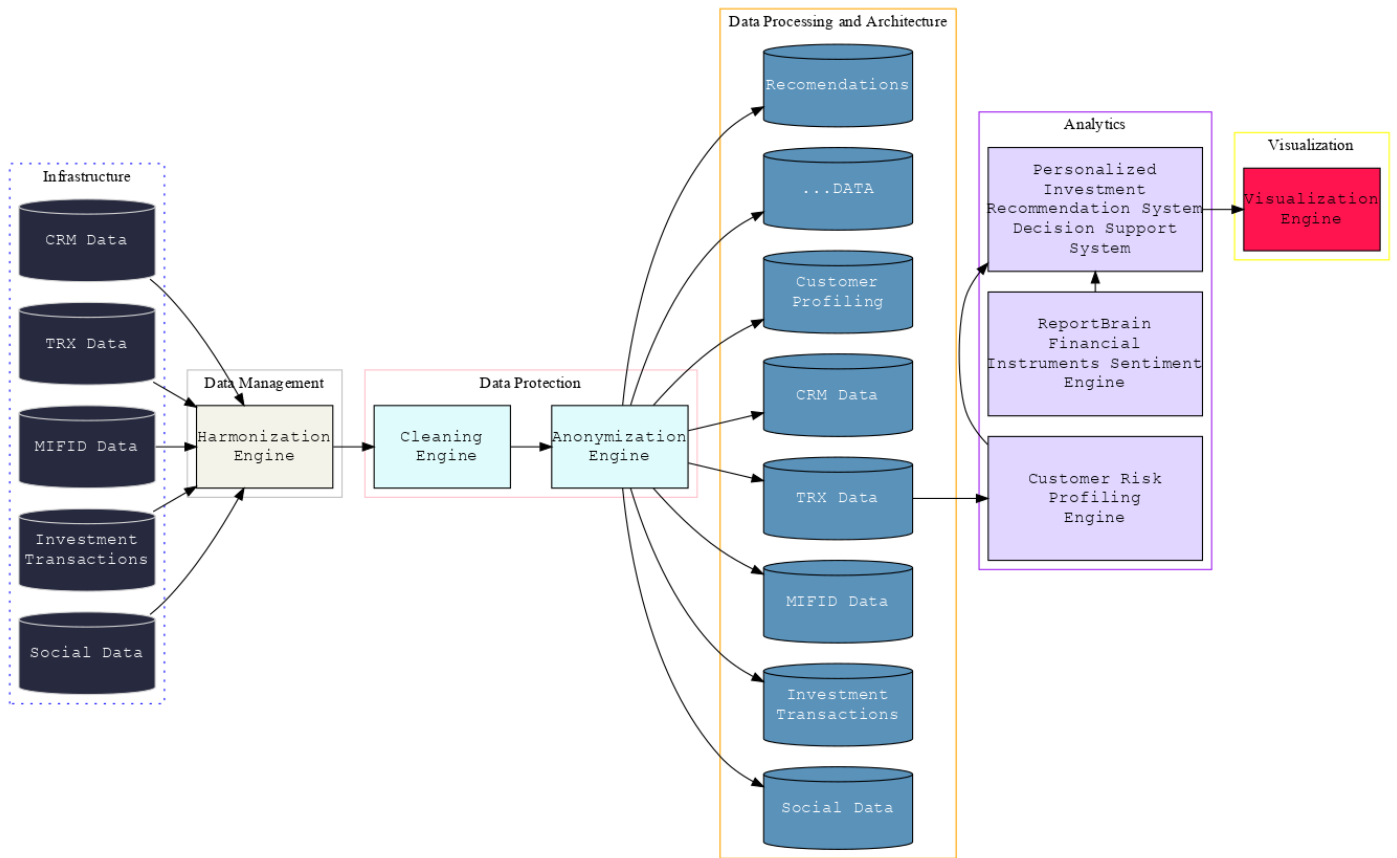


Figure 16: Pilot #6 Reference Architecture (From D2.13)

The main stakeholders for this pilot are the account officers of a bank, who will be able to provide personalized investment recommendations for customers. Recommendations based on customer (risk) profile, as well as with the relative sentiment analysis data from the news, social media, and other resources on the internet. In Pilot #6 these stakeholders will be represented by the bank, NBG(National Bank of Greece), that provides the user stories in D2.1

The configuration and roles by each partner in this pilot consists of: NBG (as Bank and Business Owner) provides customer’s data, UBI(Ubitech) process the data through the Data Management and Data Processing layer, UBI inserts this data into the datastore software provided by LXS (LeanXcale). AI algorithms (NBG), utilizing sentiment analysis data by ReportBrain(RB). University of Glasgow is now also participating enhancing AI algorithms. Finally, a final user application developed by CP (Crowdpolicy) will show the desired information and recommendations.

## 2.6.1 Technological components and Services

Going a step beyond the Pilot’s RA (Figure 16) towards the functional overview shown in (Figure 15), the High-Level Architecture presented presents the software components that build the Pilot’s use cases. This figure has been used in D6.1 to identify hardware requirements and in D2.5 to describe the technologies behind its principal components. This document links the shown software components with the corresponding RA layers, providing some details about their implementation. In this sense:

- NBG supply **raw datasets** required for the implementation of the final services. The pilot has already identified the relative customers portion of data that will be utilized, based on the existing DWH (Data Ware House).
- **Data Collection and Data Normalization components (Data Management and Protection layers of the RA):** based on Icarus from UBI, define the rules to (Data Processing layer in RA): process and harmonize, cleanse and anonymize data from NBG and insert them in a datastore available from LXS.
- **Customer Risk Profile Cluster** implemented by ML/DL Algorithm developed by NBG that cclassify customers into 4 risk profiles: Conservative, Income Seeking, Balanced, Growth Seeking. The algorithm is applied to both investors (having answered the MiFID questionnaire) and non-investors.
- **Personalized Investment Recommendation AI engine**, that will also utilize sentiment analysis data from RB, will produce the recommended instruments for investment.
- **Customer initiation and personalized recommendation** is obtained through a **visualization application** developed by CP. This application will also orchestrate the processes of analysis, initiation, execution and processing, when a new customer or new data are available.

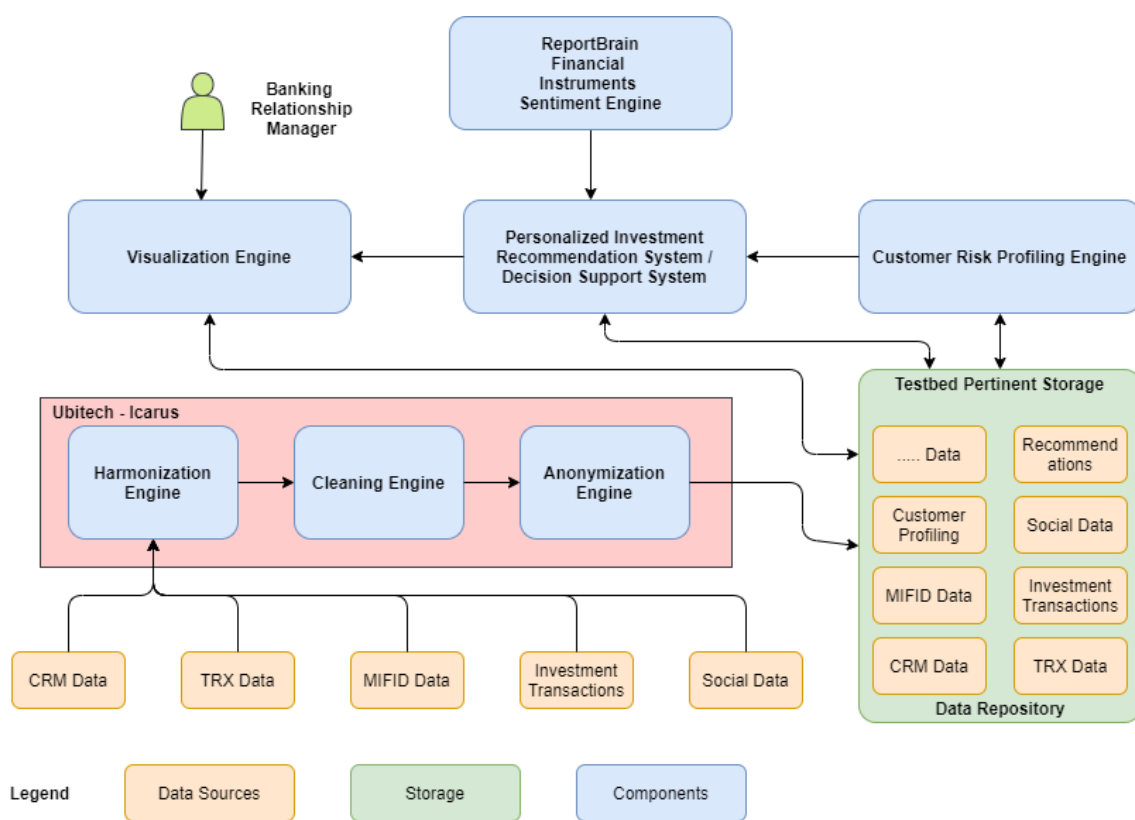


Figure 17: Pilot #6 High Level Architecture (updated From D6.1)

## 2.6.2 Data sets status

Within D2.5 in WP2 and D5.1 in WP5, the initial datasets to start Pilot #6 have been already documented and will be formally updated on their successive versions. This section presents an updated summary of the data that is put in place to start the pilot’s performance and AI initial analysis.

Data will be extracted from Operation DWH of NBG being anonymized in CSV files and using UBI Icarus will be harmonized, cleansed and extra anonymized (if needed) will be loaded to LXS Datastore.

From D5.13 Pilot #6 table and aligned with these models, the pilot manages data from NBG Operational DWH (Data Ware House) in CSV format for around 150.000 Customers:

- **Current available datasets:**

- **Deposit Account Transactions:** Data of Deposits accounts transactions for retail customers are extracted for the last two (2) years (8,91G),
- **Cards Transactions:** Data of Transactions related to Cards for retail customers for the last two (2) years (7,3GB).
- **Instruments Historical Prices:** Data for Instruments Historical Prices for the last two (2) years (0,23GB).
- **Investment Related Transactions:** Data of Investment Related Transactions for last two(2) years (0,3GB).
- **Instruments Characteristics:** Data for Instrument characteristics for matching with customers profiles (0,01GB).
- **CRM Data:** 150.000 Customers data (0,05GB) .
- **Future available datasets:**
  - Sentiment Analysis for each instrument proposed from Data Analysis as recommendation using RB information from the news or/and social media to provide to NBG customers with clearer and real-time risk results.

### 2.6.3 Testbed

Pilot’s #6 final deployment relies on MS-Azure cloud infrastructure that NBG will provided, as detailed in D6.1. Further details of the software/hardware first analysis and their results can be found in that document, but are summarised in the following table:

Component	vCPUs	RAM	Storage	Software
UBITECH Harmonization Tool	vCPUs: >= 4	>= 32GB	>= 250 GB	Java, Docker, Linux OS
UBITECH Cleansing Engine	vCPUs: >= 4	>= 32GB	>= 250 GB	Java, Docker, Linux OS
UBITECH Anonymization Engine	vCPUs: >= 4	>= 32GB	>= 250 GB	Java, Docker, Linux OS
LeanXscale Datastore (2 Data Nodes, 1 Query Data Node, 1 Metadata Node)	vCPUs: >= 4	>=8GB min >=32GB max	>=40GB	Docker, Linux OS
Visualization API & Application (2 Servers)	vCPUs: >= 4	>=16GB	>=50GB	Java Spring Boot, Angular, Express, Node JS, Docker, Linux OS
Retail Customers Risk Profiling	vCPUs: >= 4	>= 32GB.	>= 250 GB	Python Libraries
Personalized Investment Recommendations/ Decision Support System	vCPUs: >= 4	>= 32GB.	>= 250 GB	Python Libraries

Table 9 Pilot #6 Hardware/Software Requirements for the Testbed

NBG MS-Azure infrastructure it's currently deployed from NBG IT team in order to accommodate the first Proof of Concept being under development and based on the Pilot's development progress will be adjusted in terms of resources to accommodate any additional requirements related to resources.

### 2.6.4 Other non-technical requirements

Besides the technical requirements that compose the core pilot's platform, the AI technologies deployment and the data collection, we have not identified any other non-technical requirements that may affect the best outcomes for the Pilot.

### 2.6.5 Implementation of a first Proof of Concept

First Pilot #6 demonstrator (PoC) is focused on processing a subset of Cards and Deposit Accounts Transaction Data extracted from bank's Operational DWH. Figure 18 presents the functional diagram of the developed PoC.

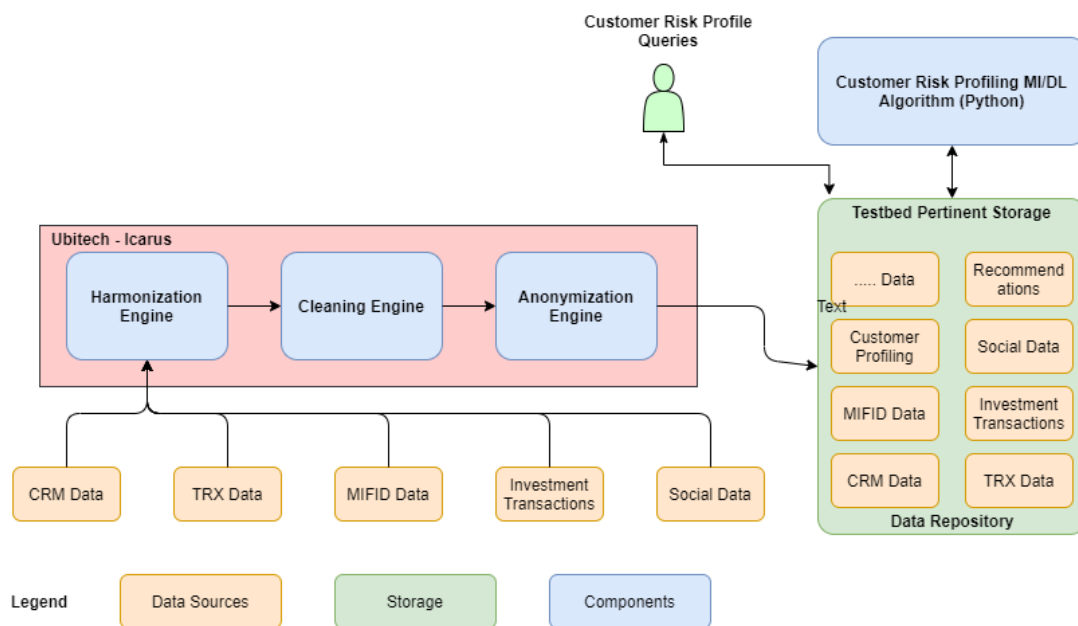


Figure 18: Pilot #6 Data collection PoC architecture

Based on the **raw datasets** available from NBG for retail customers, a first version of the relative ML/DL algorithm that will be implemented from NBG, will provide the Customer **Risk Profile clustering**. The Risk Profile will be one of the different inputs to feed the final core component: **Personalized Investment Recommendation AI engine**. The AI Engine is not available in the first PoC.

Based on the algorithm results for Customer Risk Profile clustering a web page will be provided as dashboard for visualization of the results as a way of making first demos of the PoC.

Proof of Concept execution, will provide valuable feedback for the AI approach that will work better for the Pilot execution, as well as create the common ground for the future of the development that will be required in order the full scope of the Pilot to be realised.

## 2.6.6 Next steps and timeline

Considering the PoC presented in section 2.6.5 as the current status, short-term next steps would include:

- Enhance the ML/DL Algorithm for Customer Risk Profile clustering (Feb of 2021)
- Initiation of implementation of ML/DL algorithm for personalized recommendation investment and decision support that we expect (Feb 2021)
- Modifications/Adjustments to the UBI Icarus platform Data processing software based on additional data imported from NBG Operational DWH (end of 2020)
- Design and implementation for the 1st version of relative application that will orchestrate the execution of ML/DL algorithms & visualize the results of personalized recommendations, (Feb 2021)
- Initial feed using RB Sentiment Analysis

## 2.6.7 Conclusions - Issues and barriers

Based on the work done so far, the Pilot it seems that is on track, following the implementation plan already agreed with all the contributing partners and utilize the available technology components already available or will be, as part of the INFINITECH project.

The main foreseen challenges would include:

- Implementation of the best performed ML/DL Algorithms, for this purpose we have started to evaluate some of the algorithms already available from INFINITECH partner University of Glasgow (GLA).
- Setup of the relative testbed based on the blueprint reference architecture (as will be hosted on MS-Azure)
- Calibrate the ML/DL Algorithms to provide best results for investment recommendations.

As the outcome the Bank will develop a better and more trustful relationship with its customer base, who hopefully will gradually turn exclusively select the specific bank for the entire spectrum of financial advice, products, and services. The Bank will also increase its trading volumes. The investment consultants will see their productivity improving.

## 2.7 Pilot #7 Operation Whitetail - Avoiding Financial Crime

Due to a change in pilot partners, the implementation of the pilot is subject to further specifications. During the creation of this deliverable the pilot is been redefined according to the new partners. Therefore, only a brief introduction to readiness is covered in the deliverable. Further descriptions will be included in the next version.

The goal of the Pilot is to explore more accurate, comprehensive and near real-time pictures of suspicious behaviour in Financial Crime, Fraud, and cyber-physical attacks with the final objective of stealing the bank customers' identity and money.

Within the pilot the following processes are addressed:

- KYC (Know Your Customer), for screening the vast amount of available data sources in near-real time, to ensure that KYC data is automatically updated to the most recent information available on the customer facilitating data quality.
- Customer risk profiling, based on feeding the transaction-based customer's behavioural profile data and KYC results leading to an advanced risk score that could provide a holistic customer risk profile

and will enable the business to respond quicker to newly identified risk and changes in criminal behaviour.

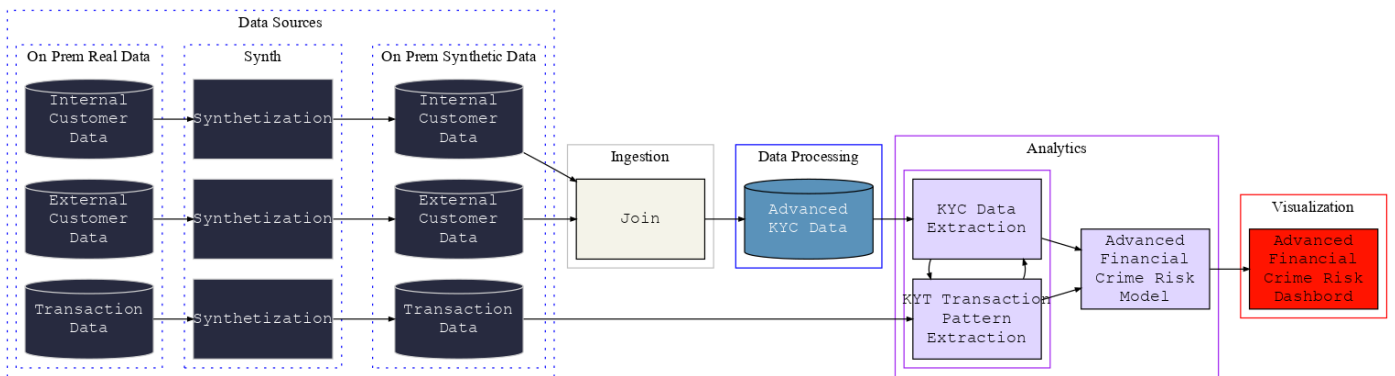


Figure 19: Pilot #7 RA (From D2.13)

Therefore, the pilot plans to utilize use synthetic or anonymized data as source. Bank internal and bank external sources of KYC data shall be joined in an advanced KYC data store.

The external data sources include public sources or sources actively shared by the customer. Information from external sources will be obtained traditionally, e.g. credit reference agencies; sanctions lists.

The advanced KYC data are used to extract a customer profile. Additionally, customer transactions patterns are extracted from the customers’ transaction data.

The workflow will produce data giving insight to the financial crime risk score. This may include a risk score, customer data, transaction patterns and details. The detailed data, which shall be produced, are yet to be specified depending on the advice of Financial Crime experts in the bank.

## 2.8 Pilot #8 Platform for AML supervision

The objective of the pilot is to develop a Platform for anti-money laundering Supervision (PAMLS), which will improve the effectiveness of the existing supervisory activities in the area of anti- money laundering and combating financing of terrorism by processing large quantity of data (Big Data) owned by the Bank of Slovenia (BOS) and other competent authorities (FIU).

The pilot will develop a platform named PAMLS that will improve the effectiveness of the existing supervisory activities in the area of ML/TF, by:

- Automated and transparent data gathering that will include data quality control.
- Improved analysis of big data coming from wide range of different sources (e.g. payment transactions, data acquired from the FI; business register etc.).
- Improved Risk Assessment (as an ongoing and cyclical process) with automated feeds from big data analysis.
- More cost-efficient risk assessment process due to less time-consuming data gathering tasks, assessments of the FI and the financial sector and semi-automated features.
- More effort-efficient risk assessment process, additional resources can be focused on the supervision of identified high risks.

PALMS will consist of four main business services:

- **Screening tool:** for screening of payment transactions and detection of potential suspicious transactions.
- **Search engine:** allowing supervisor to look for a specific transaction or a sample of transactions.
- **Risk assessment tool:** to assess the risk of the financial institution's (its inherent risks and control environment) on a risk-based approach.
- **Distribution channel:** for gathering data from other departments within Bank of Slovenia, other supervisory authorities (domestic and foreign) and the financial institutions.

## 2.8.1 Technological components and Services

Following a list of components and grouped by the Reference Architecture layers () for Data Analytics and User Interfaces.

**Data Analytics Layer** main components:

- **Risk Calculation engine** and **Complex search services**, which will be implemented specifically for Pilot8 requirements and therefore will be tailored to BOS specific:
  - Current status: 1<sup>st</sup> version developed on scrambled data
  - Next version:M27
- **Anomaly detection & prediction analysis**, which will provide functionalities for anomaly detection and prediction for time series data including Pattern analysis, which will provide analytical services on data graphs, including detection of complex patterns on data graphs:
  - Current status: to be developed
  - First version:M27
- **Stream story** is a component for the analysis of multivariate time series. It computes and visualizes a hierarchical Markov chain model which captures the qualitative behaviour of the systems' dynamics, where system is described with a group of time series.
  - Current status: to be developed
  - First version:M27

**User Interfaces Layer** main components:

- **Risk Assessments tool** – 1<sup>st</sup> version already developed, next version M27.

Figure 20 illustrates an initial logical mapping of these components to the layers approach of the INFINITECH Reference Architecture (for more details please see D2.13).

The components implemented for the first PoC include:

- **Data sources:**
  - Synthetic FI data (data about inherent risk and their control environment)
  - Implemented APIs
  - Implemented DQ 1<sup>st</sup> version
- **Risk methodology – framework**
- **Risk engine:**
  - Defined technical requirements (flexibility, scalability, DQ..)
  - 1st version implemented

- Validation & verification of risk calculations

Additionally, the Stream Story component was used to search data for meaningful patterns, however, since it was applied on scrambled data, meaningful verification of data patterns was not possible.

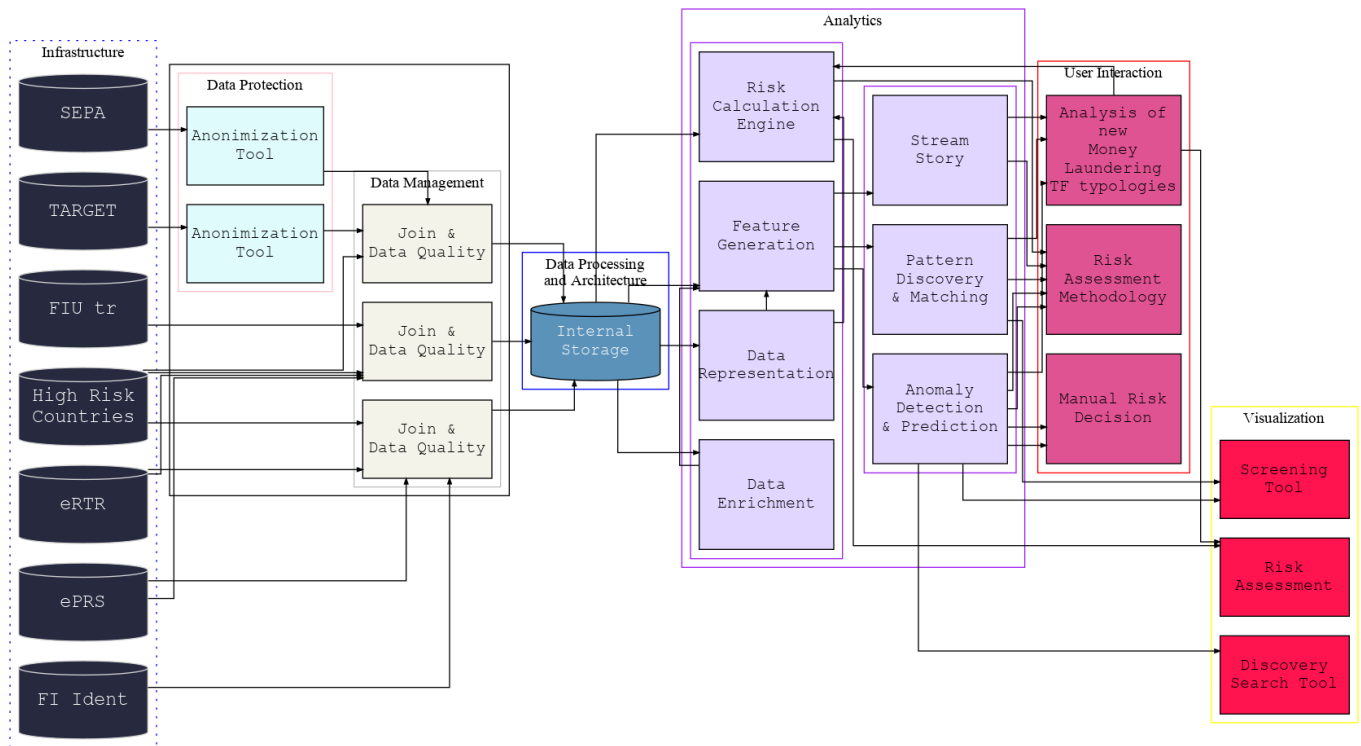


Figure 20: Pilot #8 Reference Architecture (From D2.13)

## 2.8.2 Data sets status

Relevant datasets that will be analysed within PAMLS (for more details please see deliverable D2.13):

- **TARGET2 transactions:** transactions executed by the Slovenian payment institutions - within Trans-European Automated Real-time Gross Settlement Express Transfer System;
- **SEPA transactions:** transactions executed by the Slovenian payment institutions within Single Euro Payments Area;
- **FIU transactions (public data):** transactions related to high risk countries above 15.000 EUR reported to the Slovene Financial Intelligence Unit;
- **High risk country list:** List of countries defined as high risk due to lack of or not effective AML/CTF system;
- **FI data:** information about Financial Institution (FI);
- **ePRS data:** Slovenian Business Register;
- **eRTR data:** Slovenian Transactions Accounts Register.

**Current availability of data (size, numbers, etc)**



Pilot#8 started with the development of Risk Assessment Tool for which at the moment synthetic FI data is being used. After the conformation from the BOS Governing Council real FI data will be used on the testbed developed in BOS premises, including:

- Dataset description: FI data
- Size: about 2,30 MB
- # of Records: about 45.000
- Time period: 2016 – 2019
- Data frequency: yearly

First few development phases of development of the Risk Assessment tool on FI synthetic data were done in research environment at JSI (Jožef Stefan Institute). Currently we are in the final phase of transferring it to testbed at BOS, where the actual data will be used in Risk Assessment process.

### 2.8.3 Testbed

Pilot#8 will be hosted at the Testbed on the premises of the BOS, it is ready and it has already deployed the software components and data to implement the PoC. The testbed has been specified in the following manner:

Hardware Description	
HP Z4 G4 WKS	
CPU:	Intel XeonW-2125 4.0 4C
RAM:	256GB (8x32GB) DDR4
Graphic:	NVIDIA Quadro P400 2GB (3)mDP Graphics
Disk:	Z Turbo Drv 1TB PCIe NVMe OPAL2 TLC SSD

Table 10: Testbed Specification for Pilot #8

Testbed is based on Windows operating system and include software:

- Libraries (QMiner & SNAP).
- External tools (candidates): (PostgreSQL/Elastic Search).
- Programming tools (C/C++ compilers (GNU or Microsoft, Python, Node.js)).

For more detail description of hardware and software specifications please see deliverable D6.1.

### 2.8.4 Others non-technical requirements

Due to standard security measures at BOS, in order to use the Pilot#8 testbed physical presence of JSI development team at the BOS premises is required. As a consequence of the strict measures to mitigate the spread of Covid-19 and additional security measures external parties do not have granted access to the BOS premises (JSI as a partner on Pilot#8 included). During PoC implementation the development was done on scrambled data in order to preserve data privacy. Therefore PoC implementation was done at JSI site,

however validation and initial testing was done by BOS in several phases, where risk calculations were validated. At the next phase, PoC will be transferred to the Pilot#8 testbed at BOS site. For next development phases of Pilot#8, it is crucial that during initial development of AI components appropriate test data is available, while proper validation and testing needs to be done on real data at BOS site.

## 2.8.5 Implementation of a first Proof of Concept

In accordance with the development timeline first prototype of the Risk assessment tool was developed (Figure 21):

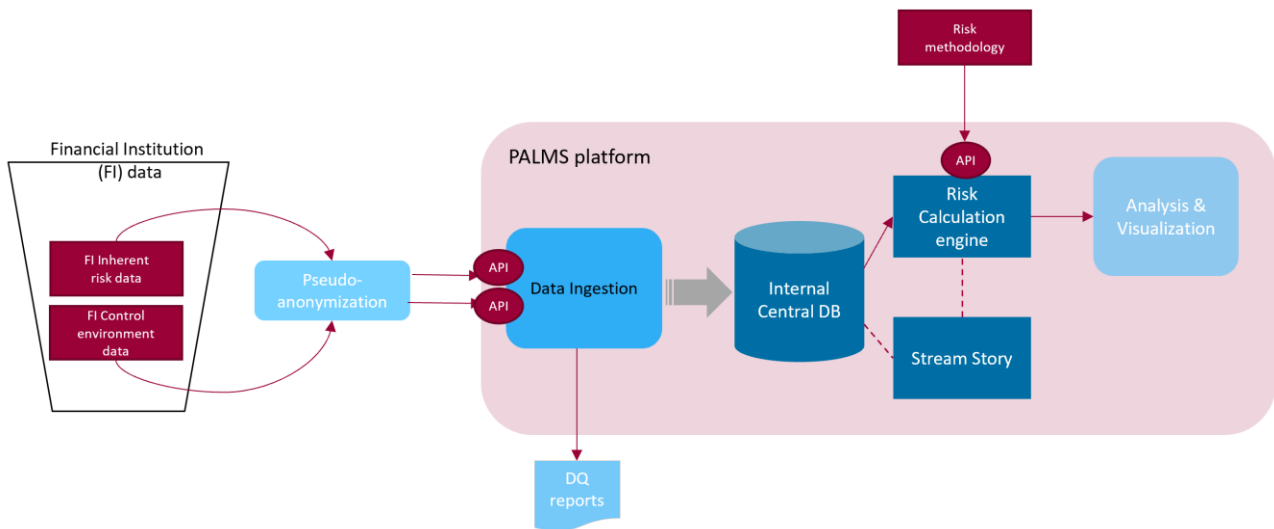


Figure 21: Pilot #8 Risk assessment tool data flow

Risk Assessment Tool – PoC was implemented in agile quick cycles. It provides the following functionalities:

**1. Sector Risk Assessment view** allows us to review the risk of all financial institutions (FIs) based on the assessment of their inherent risk and control environment in the specific year. Based on their risk, FIs are placed in the Risk Assessment Matrix in to low - medium - medium high - or high risk. Supervisory authority will focus on those presenting higher risk. Since the final risk assessment is an evaluation of inherent risk and control environment the view also enables graphical schema of those two important elements of the risk assessment.

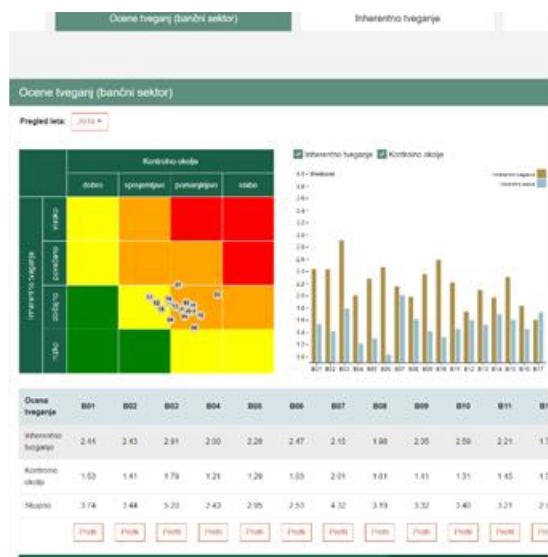


Figure 22: Sector risk Assessment View

Changes in the risks of the specific FI is also an important factor. Therefore Sector Risk Assessment view enables also historical view for selected FIs.

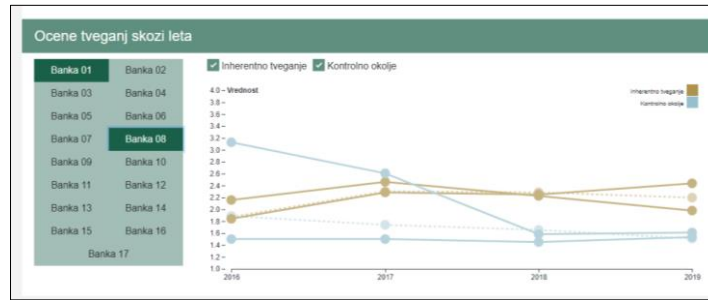


Figure 23: Sector Risk Assessment view- historical view

**2. Inherent risk / Control environment view:** Inherent risk (and similar for control environment) consists from different risk areas and those consist from different elements. In this view supervisor can drill down to the specific elements and to compare FIs amongst each other. Also supervisor receives information which areas or elements of the inherent risk or of the control environment present more risk for a specific FI and can therefore focus on those areas during the on-site supervision.

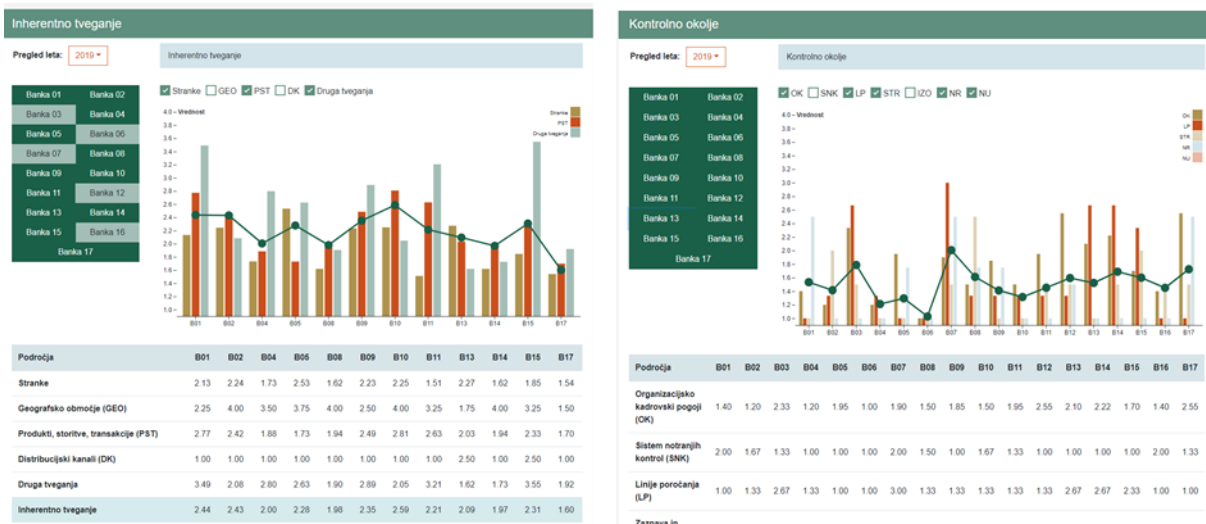


Figure 24: Inherent risk and control environment view

**3. Bank Profile view** enables the supervisor to select a FI for a detailed review. In the first version of PoC the view consists from FI basic information (FI ID Card), graph on the FIs risk assessment changes through the year (historical view) and detailed information on FI inherent risk and control environment.

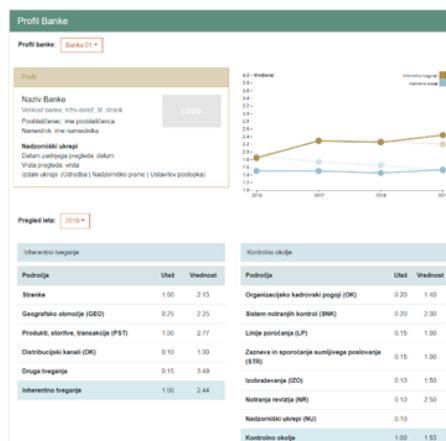


Figure 25: Bank Profile View



scalable transaction graph analysis system is being developed that runs on HPC cluster and that can process the growing transaction graph without encountering performance bottlenecks.

The main innovation of the pilot lies in **the applicability of HPC technologies to analyse Blockchain (huge) transaction graphs**, to quickly detect possible frauds based on blacklists.

Following the main components of the pilots and the partners in charge of the development:

- Blockchain Transaction Dataset Preparation Component (developed by BOUN (Bogazici University)),
- Scalable Transaction Graph Analysis Component (developed by BOUN (Bogazici University)),
- User Interface for Blockchain Transaction Reports and Visualization Component (developed by AKTIF Bank).

The final users will be banks who need to do analysis of blockchain addresses. Developed services can also be offered as a service to companies who need to do such checks, for example, companies that accept token payments.

The first year aimed at massive blockchain dataset preparation, an HPC based cluster parallel transaction graph analysis system construction and coding of traversal based graph algorithms. The second year will concentrate on machine learning based approaches for analysis using, in particular, the graph system developed in the first year for feature extraction.

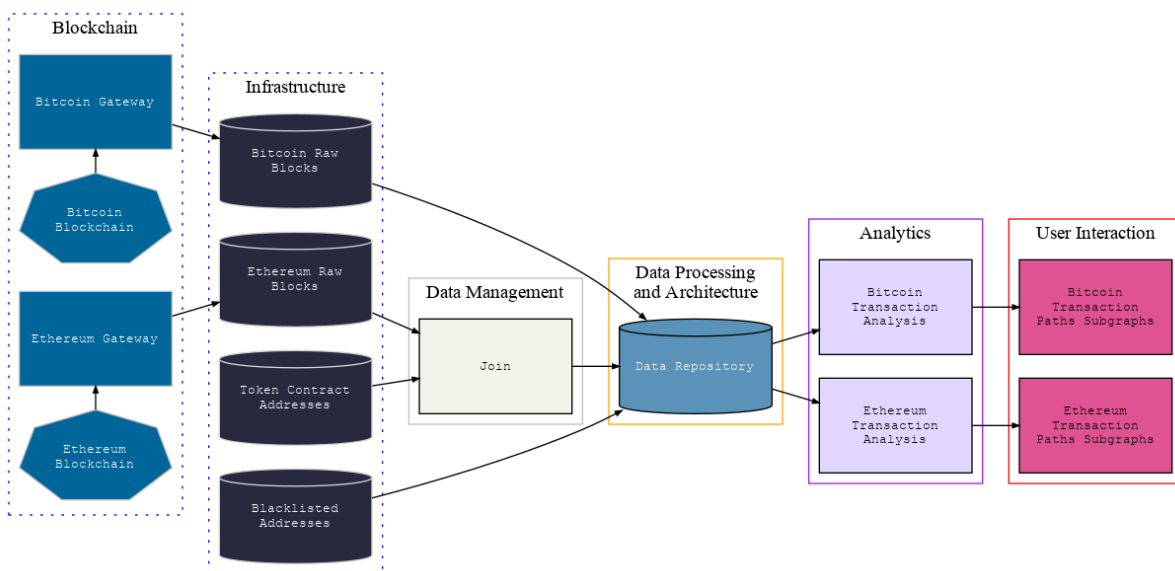


Figure 27: Pilot #9 Reference Architecture

Following a list of blocks and how these are related to the INFINITECH Reference Architecture. More details of the Pilot #9 RA in deliverable D2.13.

**Platform for data gathering** (Related Reference Architecture Layers (Figure 27): Blockchain, Infrastructure and Data management).

Blockchain dataset component is implemented as scripts that retrieve blockchain data as raw block data and parse these to extract crypto-currency and token transaction. Sources of blockchain raw data are Cloudflare Ethereum Gateway, Google Bigtables and blockchain nodes.

**Big Data management** (Related Reference Architecture Layers (Figure 27): Data Processing and Analytics).

One cannot assume that massive blockchain data will fit in one computer node. Therefore, a distributed in-memory storage on an HPC cluster is essential. Currently, Scalable Transaction Graph Analysis Component

which is implemented using C/C++ and MPI message passing libraries constructs a partitioned graph in parallel and provides big data management and processing capability.

**Statistics, analysis, AI (Related Reference Architecture Layers (Figure 27): User interface).**

In order to carry out analytics and report various statistics, two types of approaches are to be utilized (i) Graph Algorithms Approach and (ii) Machine Learning Approach. For Machine learning, K-Means, Support Vector Machines, Naive Bayes, Logistic Regression, Random Forest, Artificial Neural Networks (Multilayer Perceptron) methods will be used by making use of the existing Python Scikit-learn and Pytorch machine learning software. Analytics layer, in Figure 27:, shows these functionalities.

**Readiness, matureness, level of development (TRL level)**

Currently, a proof of concept (PoC) implementation of the pilot is available. As a whole, the current level of development is at TRL3. On the other hand, industrially relevant environment for Pilot9 is defined to be an environment where the real world blockchain data is used. When carrying out our tests in Pilot9, we do use massive industrially relevant blockchain data. The eventual target TRL level is TRL7.

## 2.9.1 Data sets status

Pilot 9 will make use of public Ethereum and Bitcoin blockchain data. We have extracted transactions from these blockchains since their deployment. Table 11 and Table 12 list various information about the Ethereum and Bitcoin blockchain datasets respectively. Various information such as the total numbers blocks, transactions and addresses, symbols of token contracts tracked and data sizes are reported in the tables.

Blocks	0 - 9499999
Time Coverage of Blocks	30.07.2015-17.02.2020
No. of Transactions	633 762 485
No. of Addresses	69 223 762
No. of 31 Major ERC20 Token Transfer Transactions:	24 646 152
List of symbols of 31 Major ERC20 Tokens	USDT PAX EURS GUSD TRYB BAT CEO LNK HT HEDG MKR CRO VEN INO INB SNX MOF ZRX SXP OKB XIN SAI HOT DAI HPT BUSD XAUT USDC SUSD HOG QCAD
Bz2 zipped size (extracted data)	18GB
Uncompressed size (extracted data)	70 GB
Dataset Repository link	<a href="https://zenodo.org/record/3669937#.X6OZe3gzY1I">https://zenodo.org/record/3669937#.X6OZe3gzY1I</a>

Table 11 Ethereum Mainnet Blockchain Dataset Details

Blocks	0 - 623467
Time Coverage of Blocks	03.01.2009 - 29.03.2020
No. of transactions	516.305.390

Bz2 zipped size	91 GB
Uncompressed size	310 GB

Table 12 Bitcoin Blockchain Dataset Details

## 2.9.2 Testbed

Figure 28 depicts the testbed which is currently set-up and running on the Amazon cloud. The following is the hardware and software configuration that is used for the testbed:

Hardware:

- HPC Cluster on Amazon Cloud (16 c5.4xlarge instances), each instance having 16 virtual CPUs, 32 GiB memory and 500 GB SSD storage.
- A medium Amazon instance for running message queue.

Software:

- Ubuntu Linux operating system
- StarCluster HPC cluster toolkit.
- MPI message passing interface
- Rabbit MQ message queue
- Metis Parallel graph partitioner
- Vis.js open source graph visualization software for web interface.

## 2.9.3 Other non-technical requirements

There are no other non-technical requirements at the moment.

## 2.9.4 Implementation of a first Proof of Concept

Figure 28 shows the architecture of the Proof of Concept system that has been implemented.

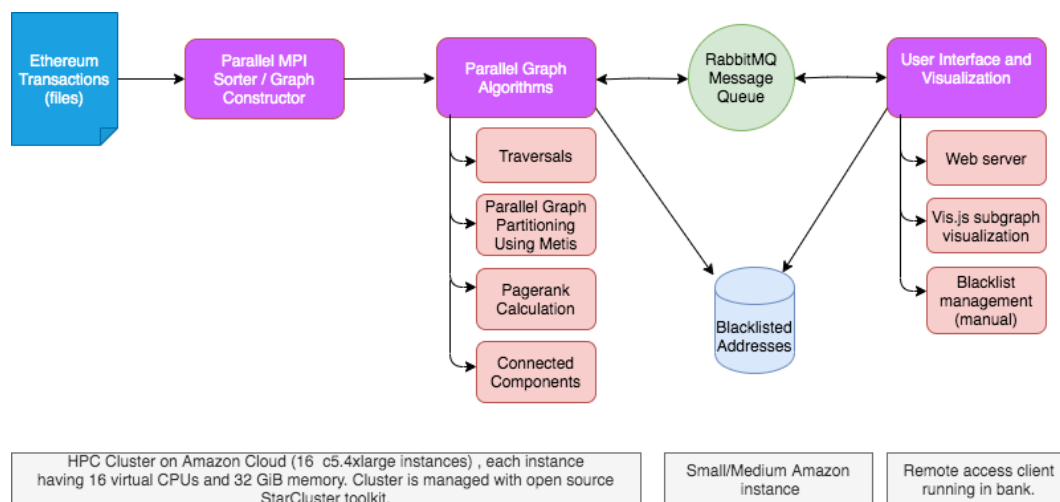


Figure 28: Pilot #9 Architecture of the Proof of Concept

### Services to be implemented according the user stories

PoC currently offers parallel scalable blockchain transaction graph construction, parallel graph traversals that trace customer addresses to blacklisted addresses by returning the traced subgraph. Parallel Pagerank algorithm that finds important addresses is also offered as a service. The transaction graph can also be partitioned in parallel using the Metis software. These services are offered on the whole dataset graph having 633M transactions.

### Components implemented, interactions and deployment

The following components of the project have been built as proof of concept:

1. Blockchain Transaction Dataset Preparation Component
2. Scalable Transaction Graph Analysis Component
3. User Interface for Blockchain Transaction Reports and Visualization Component

Component 1) parses Ethereum raw data to extract transactions which are saved as files. 2) constructs the distributed and partitioned graph on HPC cluster using the transaction files and performs parallel graph algorithms. 3) communicates with 2) via RabbitMQ message service, submits queries and displays returned results on web page and produces graph visualization output using the Vis.js package.

### Data sets available (kind of data, size, numbers)

PoC currently uses only the Ethereum blockchain dataset whose details are listed in Table 11.

### Testbed and infrastructure

Figure 28 illustrates the testbed and the infrastructure of the first Proof of Concept which has been implemented.

## 2.9.5 Next steps and timeline

The first year concentrated on setting up our massive Ethereum Mainnet dataset and parallel graph analysis infrastructure. The second year will concentrate on:

- The use of Python Scikit-learn and Pytorch machine learning software's in order to analyse the subgraphs that are returned from graph analysis system and report score that indicate degree of illicitness. K-Means, Support Vector Machines, Naive Bayes, Logistic Regression, Random Forest, Artificial Neural Networks (Multilayer Perceptron) methods will be explored.
- Parallel feature extraction will be carried out on massive blockchain data using the parallel graph analysis system and these features will be fed to machine learning algorithms.
- Performing graph analysis on Bitcoin transaction graph. Whereas Ethereum is account based, Bitcoin is unspent output (UTXO) based and hence their transaction graph structures differ. We will update our current distributed graph structures to support many-to-many type (directed hyperedge) transactions that are possible in Bitcoin.
- Development of more efficient parallel connected components algorithm.
- Further developments of web interface and graph visualization system in order to provide an easy to use interface that will interact with the graph analysis system.
- Implement blacklisted addresses database management interface.



## 2.9.6 Conclusions - Issues and Barriers

The first year of the Pilot9 has focused on (i) collection and parsing of public massive blockchain data and (ii) design and development of a scalable parallel transaction graph system (iii) development of a simple web interface that would query the graph system and output visualizations of subgraphs returned. We concentrated mainly on Ethereum Mainnet blockchain data, because it was more challenging to deal with due to smart contract support. Code needed to be written to extract transactions from token contract calls.

Whereas implemented parallel algorithms for graph construction, Pagerank computation, tracing and extracting of subgraphs have been tested successfully, our parallel connected algorithms has an issue with it. It is working on small test graph with 1M transactions. But on the whole 633M transaction graph, it is taking too long and not possibly terminating either due to a bug or because the parallel algorithm coded is not efficient due to excessive communication. This will be fixed in the future by coding a more efficient algorithm.

Even though there exists massive public blockchain transaction data and this data can be obtained easily by writing scripts, the same cannot be said for blacklisted addresses. Publicly available blacklisted addresses had to be located through google searches by hand and extracted manually. Collection and tagging of blacklisted addresses information remain as challenging issues because often this type of data may be private and not publicly available.

For machine learning, we need data that can be used for training in our models. In particular, licitness and illicitness information about addresses are needed, but little information is available about this – just the roughly 4K Ethereum blacklisted addresses available from various sites on the web are available to start with. On the other hand, there are roughly 70 million addresses on the Ethereum Mainnet. Hence availability of illicit addresses is limited. Identities of owners of addresses are also not available. This is currently the biggest issue and barrier that we currently have. However, the fact that a parallel cluster graph analysis system has been built means that we can do fast graph queries and traversals on massive data. As a result, we plan to tackle these challenging issues and barriers, by developing graph algorithms that provide information about licitness and illicitness. For example, Pagerank algorithm can be ran to find out important addresses. These addresses are more on the side of licitness since they are addresses of popular services like exchanges that are regulated. Since exchanges verify addresses of customers, then transactions going to addresses from such services are more likely to be licit since KYC/AML checks are carried out by exchanges. Hence, the graph traversal algorithms can be used to report features related to possible licit or illicit addresses in this manner without actually having information about the addresses in question. These extracted features can then be used in Machine Learning algorithms.

## 2.10 Pilot #10 Real-time cybersecurity analytics on financial transactions' data

Pilot #10 aims to significantly improve the detection of cases of suspected fraudulent transactions, to enable the identification of security-related anomalies while they are occurring by the analysis in real-time of the financial transactions of a home and mobile banking system. The ability to detect anomalies faster (i.e. in real time) and to unveil potential hidden patterns of cyber-attacks are among the main innovations of the pilot.

The use case envisages a pre-processing of transaction data and model training in a batch layer (to periodically retrain the predictive model with new data) while in a stream layer, the real time fraud detection is handled based on new input transaction data.

Figure 29 shows a logical view of the components identified for the Pilot#10 according to the mapping with the INFINITECH Reference Architecture (please see D2.13 for more details). The list of main components to be deployed and used in the Pilot#10 includes:

- **Identity Management System:** It is a cross cutting system that guarantees user authentication, authorization and management. It allows or denies the access to the federated services that run within the architecture.
- **Role Management:** It implements and handles the roles and the privileges that can be associated to the users. It is often tightly coupled with the Identity Management System.
- **Message Broker:** works as an intermediary software that allows system components to communicate each other effectively, implementing a common communication protocol over message buses.
- **Resource Manager:** It is a lower-level software that consents to handle and use the infrastructure resources seamlessly and dynamically, according to the number of requests received per time interval.
- **Pseudoanonymizer:** tool to pseudonymize personal or sensitive data at source, in order to preserve privacy according to GDPR regulation
- **Filter:** Filtering component to remove specific rows and columns
- **Join:** Service to join two or more datasets where at least one column must be the same
- **OneHotEncoder:** Service to transform categorical variables into numerical ones
- **Clustering (Kmeans):** Given a set of observations  $(x_1, x_2, \dots, x_n)$ , where each observation is a  $d$ -dimensional real vector,  $k$ -means clustering aims to partition the  $n$  observations into  $k$  ( $k \leq n$ ) sets  $S = \{S_1, S_2, \dots, S_k\}$  so as to minimize the within-cluster variance.
- **Random Forest:** An ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set
- **Results:** Storage that contains all the processed data elaborated by the workflow.
- **Visualization:** The service that gathers the resulting datasets to be delivered to the visualization clients.

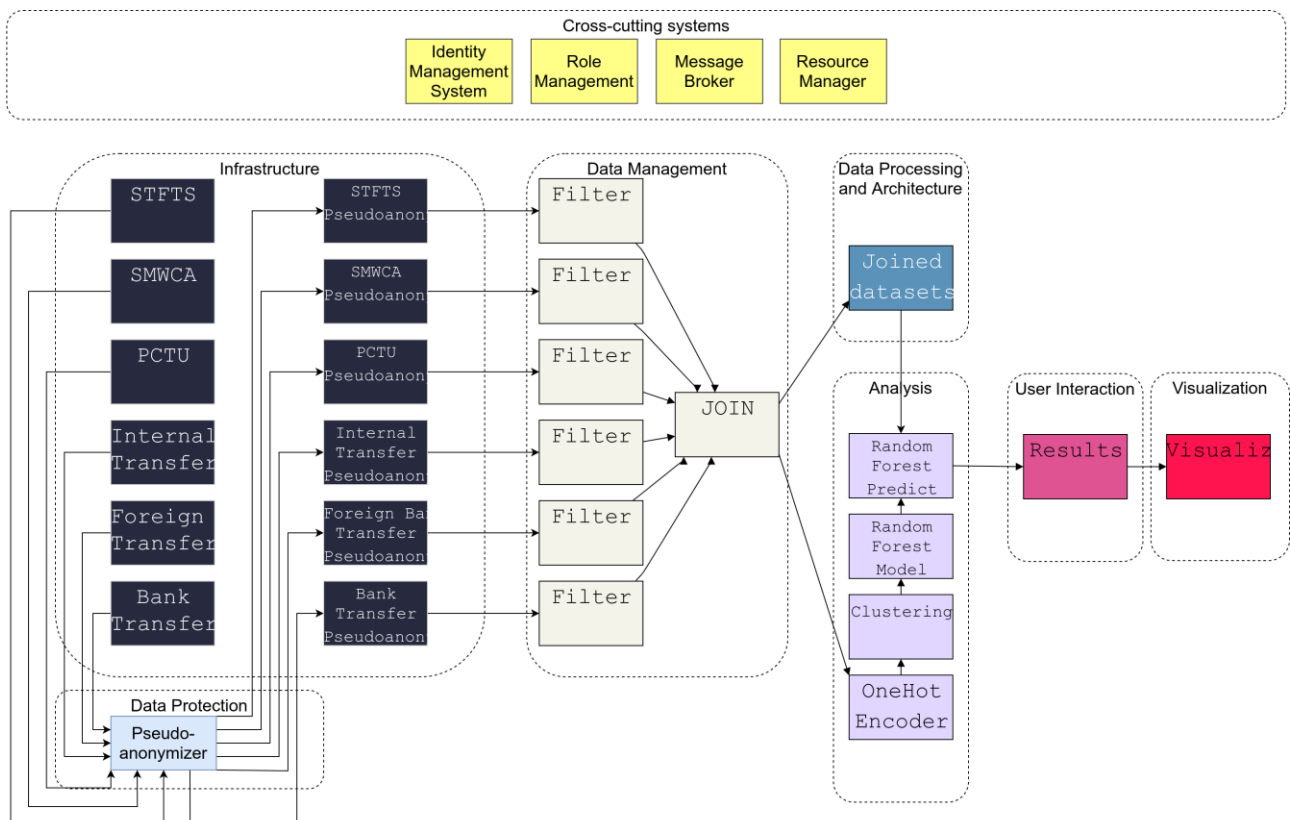


Figure 29: Pilot #10 Reference Architecture

For the Pilot #10 aims, ALIDA (<https://home.alidalab.it/>) is adopted and extended to design Big Data Analytics (BDA) services batch and stream workflows. In a nutshell, ALIDA is a micro-service based platform, developed by ENG (Engineering), for composition, deployment, execution and monitoring of workflows of BDA services; it is entirely developed with open source technologies.

ALIDA offers a catalogue of BDA services (for ingestion, preparation, analysis, visualization), implemented as Spring Boot Applications and deployable as docker images. User designs his own (stream/batch) workflow by choosing the BDA services from it, indicates which Big Data set he wants to process, launches and monitors the execution of the workflow and personalizes the results visualization by choosing from a set of available graphs. All this without worrying about having software developer skills or particular knowledge on big data technologies.

Some BDA services for preparation and machine learning, as KMeans and Random Forest modelling and prediction, are already available within the ALIDA Catalogue. Even though they need to be reviewed (and in some cases redesigned) to meet specific pilot requirements.

Concerning the remaining BDA services (especially pseudo anonymization one) the pilot will make use of the services made available within the project.

Preliminary step:

To load data sets related to several types of transactions (SEPA bank transfer, foreign bank transfers, internal transfers of funds, PCTU, SMWCA, STFTS) into the HDFS storage of the ALIDA instance, by means an ingestion job.

Batch processing, building and labelling clusters (training):

Stored data sets are properly filtered (to remove some columns and rows unneeded for the ML) and joined to get only one unlabeled data set to be used for the unsupervised machine learning.

In this phase the goal is to cluster such data, to create labeled samples to feed the supervised machine learning classifier of the next phase. Clustering process groups the data according to automatically detected similarities. These clusters/groups still need a domain expert, PI (Posteltaliane), who determine which clusters present a fraudulent behaviour and properly assign labels to such clusters.

Stream processing:

After learning the mapping, the Random Forest (RF) classifier can map new real-time unlabeled transaction data to their corresponding high-level information (i.e. label) on the basis of the model trained in the batch layer. In that way, financial fraud events can be detected while happening.

**Readiness, matureness, level of development (TRL level)**

ALIDA will be enhanced with the design and deployment of analytics services for Pilot #10.

Furthermore the deployment of the bundle of BDA services workflows, will be made semi-automatic so to be properly integrated in on-premise client systems.

TRL Present: TRL 5

Target: TRL 7

In the following table, for each service involved in the Reference Architecture (Figure 29), the readiness is indicated

Name	Description	Level of development
------	-------------	----------------------

<b>Identity Management System (IDM)</b>	It is a cross cutting system that guarantees user authentication, authorization and management. It allows or denies the access to the federated services that run within the architecture.	Ready (meaning the third-party IDM integrated into ALIDA access management)
<b>Role Management (RoM)</b>	It implements and handles the roles and the privileges that can be associated to the users. It is often tightly coupled with the Identity Management System.	Ready (meaning the third-party RoM integrated into ALIDA access management)
<b>Message Broker (MB)</b>	It works as an intermediary software that allows system components to communicate each other effectively, implementing a common communication protocol over message buses.	Ready (meaning the third-party MB integrated into ALIDA access management)
<b>Resource Manager (ReM)</b>	It is a lower level software that consents to handle and use the infrastructure resources seamlessly and dynamically, according to the number of requests received per time interval.	Ready (meaning the third-party ReM integrated into ALIDA access management)
<b>Pseudoanonymizer</b>	Tool to pseudonymize personal or sensitive data at source, in order to preserve privacy according to GDPR regulation	To be developed
<b>Filter</b>	Filtering component to remove specific rows and columns	Developed and registered within the ALIDA catalogue. Updates to be assessed according to the pilot needs)
<b>Join</b>	Service to join two or more datasets where at least one column must be the same	Developed and registered within the ALIDA catalogue. Updates to be assessed according to the pilot needs
<b>StringIndexer and OneHotEncoder</b>	Service to transform categorical variables into numerical ones	Developed and registered within the ALIDA catalogue. Updates to be assessed according to the pilot needs)
<b>Clustering (Kmeans)</b>	Given a set of observations $(x_1, x_2, \dots, x_n)$ , where each observation is a $d$ -dimensional real vector, $k$ -means clustering aims to partition the $n$ observations into $k$ ( $\leq n$ ) sets $S = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster variance	Developed and registered within the ALIDA catalogue
<b>Random Forest</b>	An ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set	Developed and registered within the ALIDA catalogue

<b>Visualization</b>	The service that gathers the resulting datasets to be delivered to the visualization client	To be developed
----------------------	---	-----------------

Table 13 Pilot #10 Components Readiness

## 2.10.1 Data sets status

The data sets in input to the BDVA workflow involved in Pilot #10 are related to several types of transactions:

- Bank Transfer SEPA (The Single Euro Payments Area (SEPA): a payment-integration initiative of the European Union for simplification of bank transfers denominated in euro. SEPA covers predominantly normal bank transfers.
- Foreign Bank Transfers: international transfer of funds between banking accounts held in financial institutions located in different countries
- Internal transfers of funds: transactions involving transfer of funds from the Bank current account of one payer to that of another Bank account’s holder
- PCTU: Top-ups phone credit
- SMWCA (Sending Money to people Who do not have a Current Account): is a means of payment for transferring money to those not holding a bank account
- STFTS (Secure telematic fund transmission system) Neither the sender nor the receiver of the money being transferred need to hold a banking account

For each data set, when appropriate, data as name, date of the operation, IBAN of the beneficiary will be considered. The datasets that will be used in Pilot #10 are Fully Synthetic ones and they should be considered very realistic data as they are built using the same statistical distributions extrapolated from real data flows.

### Current availability of data (size, numbers, etc)

Currently, the available data is a large dataset containing “Bank Transfer SEPA” transactions.

Dataset description:

- Size: about 210 MB on NTFS filesystem
- # of Records: about 1 Million (Bank Transfer SEPA transactions)
- Time period: 1st January 2020 – 30th September 2020
- Users generating transactions: about 10.000
- Includes frauds: Yes
- Marked frauds: No
- Number of included money mules: about 100

Such data set was already ingested within ALIDA cluster by means of a job that transfers them from an SFTP server to ALIDA HDFS.

Thus, the data ingestion process is currently active and working. The dataset is visible and available for the analysis process.

## 2.10.2 Testbed

With regards to the pilot #10 design and execution, the testbed definition (that is the setting of hardware resources, like Storage, Compute and Network...) aims to consider the deployment of an instance of ALIDA asset.

The set of resources needed is described in the following picture:

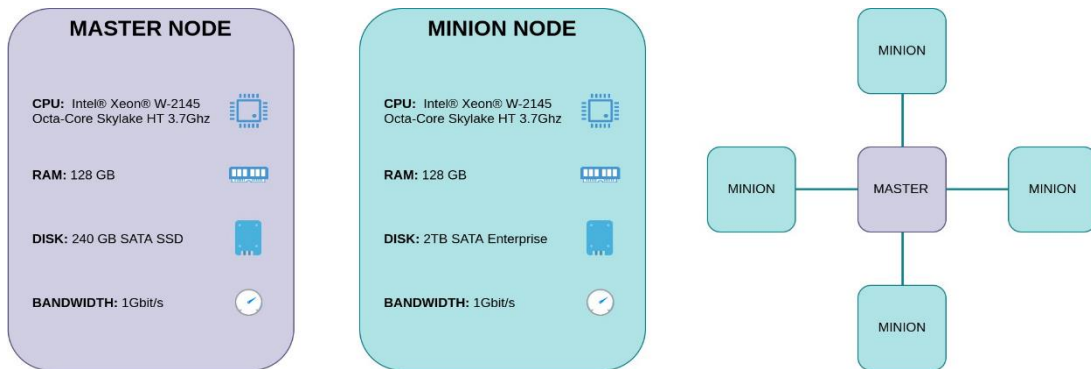


Figure 30: Pilot #10 Testbed logic

The infrastructure setup consists in a single-master four-worker nodes running a as kubernetes on-premise cluster, each node has a wide enough set of allocable resources to run the testbed safely and without running into disk pressure and memory pressure issues for the expected workload. In any case, the system can be scaled both horizontally and vertically. The machines are equipped with 128 GB of RAM, 2TB of storage and one octa-core 3.7Ghz processor.

ALIDA is cloud native software, this means that can be seamlessly deployed both in an on-premise environments and on the cloud environments provisioned by the widely known providers such as Microsoft Azure, Amazon AWS and Google Cloud Platform.

To summarize, the software requirement to get ALIDA up and running are:

- Helm and Tiller 2.14+.
- Kubernetes 1.14+.
- To enable ingresses, a valid ingress provider is required, Traefik is recommended.
- A DNS service provider is recommended to use ingresses with Traefik.
- A persistent volume provisioner support in the underlying infrastructure.

### 2.10.3 Next steps and timeline

Next development steps and timeline to implement the user story and components are:

- Data ingestion of other transaction data sets.
- Improvement of the mechanism to make a data synthetic.
- Data model design for data preparation so to apply clustering algorithm.
- Development of BDA services.
- Stream layer development.
- Batch layer to be completed.
- Integration of workflows with missing services (see cross cutting group in the figure, UI ...).
- Need to use Pseudoanonymization service to be evaluated (since all data sets are synthetic).

### 2.10.4 Implementation of a first Proof of Concept

Current implementation status on Pilot#10 is shown in Figure 31.

PI create Synthetic and Realistic data set on “Bank Transfer SEPA” transactions that are consistent with the real data present in the data operations environment. These data sets are going to be used by Pilot #10 and, more in concrete, for the first PoC. To develop the services and workflows and ALIDA instance was deployed on ENG premise. As a Preliminary step: a job to transfer synthetic data set on “Bank Transfer SEPA” transactions from an SFTP server to ALIDA HDFS, was designed and it is up and running.

With the data ready to be processed, and using ALIDA, a first Batch processing/workflow has been created. This workflow converts qualitative fields into quantitative one, train a KMeans model and makes the clustering process. The Figure 31:shows developed ALIDA workflow based on three steps (string-indexer, trains the data with a KMeans models and the clustering creation).

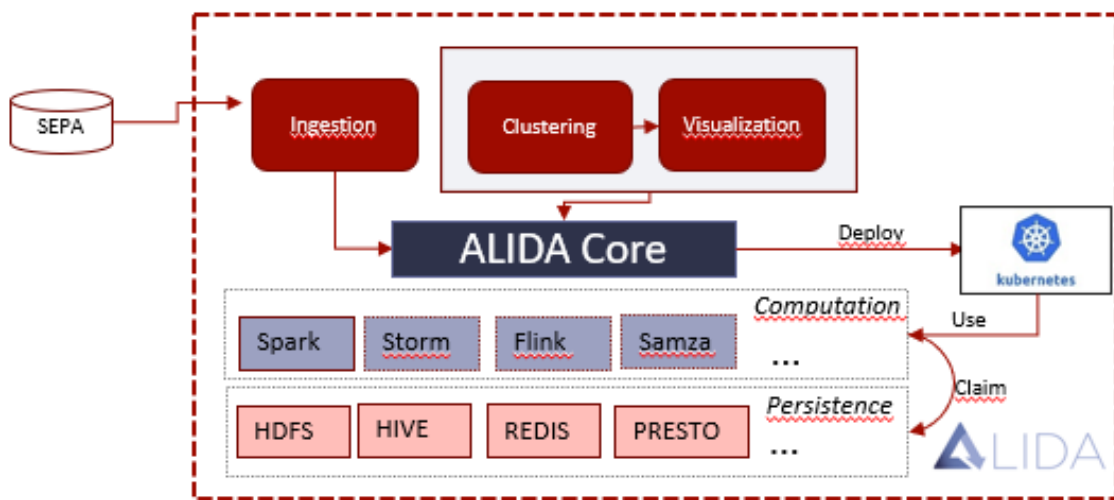


Figure 31: Pilot #10 PoC (October 2020)

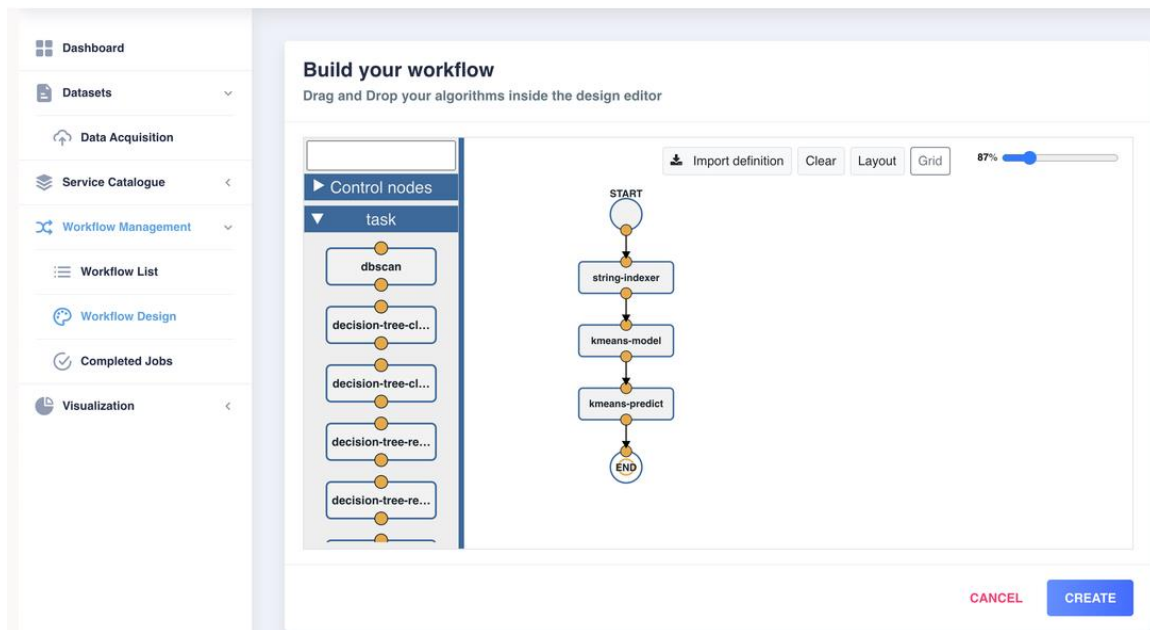


Figure 32: Pilot #10 Clustering workflow

After that, the data is grouped and visualized by clusters (Figure 33:). Here a domain expert has to label which clusters would be suspicious of fraud. After that the Stream processing would start labelling and detecting new incoming data in real time. But this part is not implemented yet.

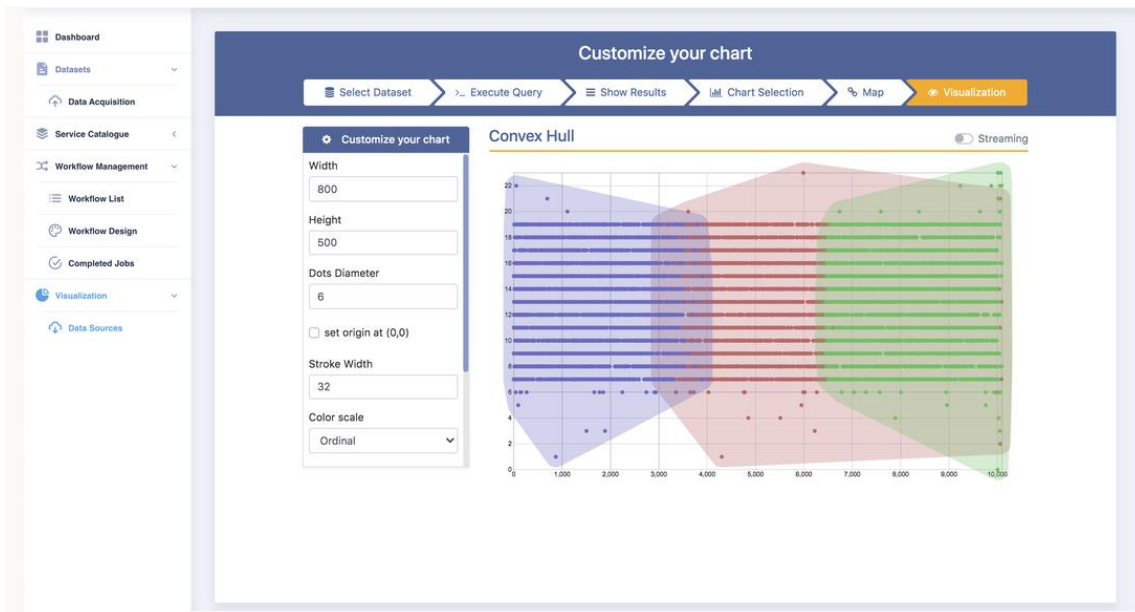


Figure 33: Pilot #10 Clustering Results

### 2.10.5 Conclusions - Issues and Barriers

Current setup clearly demonstrates some of the most relevant capabilities of the pilot:

- availability of a significant dataset for analysis
- data collection from source and preparation
- data ingestion
- AI model training
- prediction
- data visualization

Some pre-processing services are supposed to be needed in both batch and stream stages before the ML algorithms are invoked. They will be implemented once the data schema on transactions will be defined.

In order to fulfil GDPR requirements both at pilot stage and in potential production stage with real transaction data, a fully synthetic dataset will be also pseudonymized at source. Currently, synthetic data are pseudonymized at generation time, therefore data analysis will work on pseudonymized data, but we expect a pseudonymization tool will be made available in the framework of the INFINITECH project for potential production use with real data.

#### Expected Business Impact of Technologies adopted in this Pilot

Frauds on financial services are an ever-increasing phenomena and cybercrime generates multi-million revenues, therefore even a small improvement in fraud detection rates would generate significant savings.



This viewpoint, built on information sharing activities currently running in the banking sector, is also reinforced and strengthened by trusted industry reports.<sup>8910</sup> With some surveys and reports pointing to issues, such as: “recover less than 25 percent of fraud losses”, “Increase fraud typologies globally, from recent years, include identity theft and account takeover, cyber-attack, card not present fraud and authorized push payments scams”, “6 is the average number of frauds reported per company studied”, “56% asked companies conducted an investigation into their worst fraud incident. many organisations are failing to respond effectively”. These, and other issues in these reports, demonstrate the importance of developing new technologies and approaches, such as real time analytics, to enhance the need of fighting against cyber frauds.

## 2.11 Pilot #11. Personalized insurance products based on IoT connected vehicles.

This pilot focuses on car insurance and risk analysis by developing two AI powered services: Pay as you Drive, that allows the insurance company to adapt prices by classifying the driver according the way he/she drives; and the Fraud Detection which helps to identify the actual driver of a vehicle involved in an incident. These two services rely on a driving profiling tool that requires datasets from connected vehicles to define, identify and train the different profiles as ML models. Other external data sources, such as traffic incidents or weather, will be used to classify the driver, contextualizing its assigned driving profile. An overview of Pilot #11 is given Figure 34.

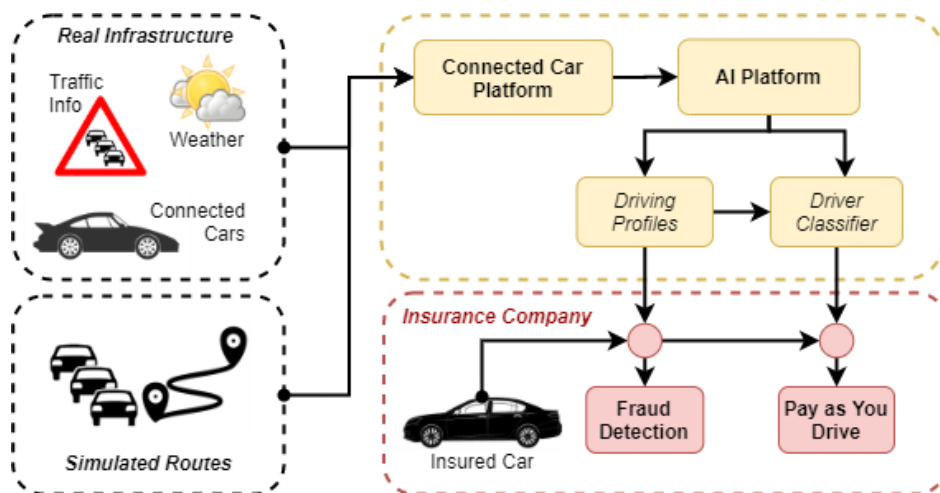


Figure 34: Pilot #11 Personalized insurance products based on IoT connected vehicles overview

<sup>8</sup> KPMG, “The multi-faceted threat of fraud: are banks up to the challenge?”, <https://home.kpmg/xx/en/home/insights/2019/05/the-multi-faceted-threat-of-fraud-are-banks-up-to-the-challenge-fs.html>

<sup>9</sup> UK Finance, 2020 Half year fraud report, <https://www.ukfinance.org.uk/policy-and-guidance/reports-publications/2020-half-year-fraud-report>

<sup>10</sup> PwC, Global Economic Crime and Fraud Survey 2020, <https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html>

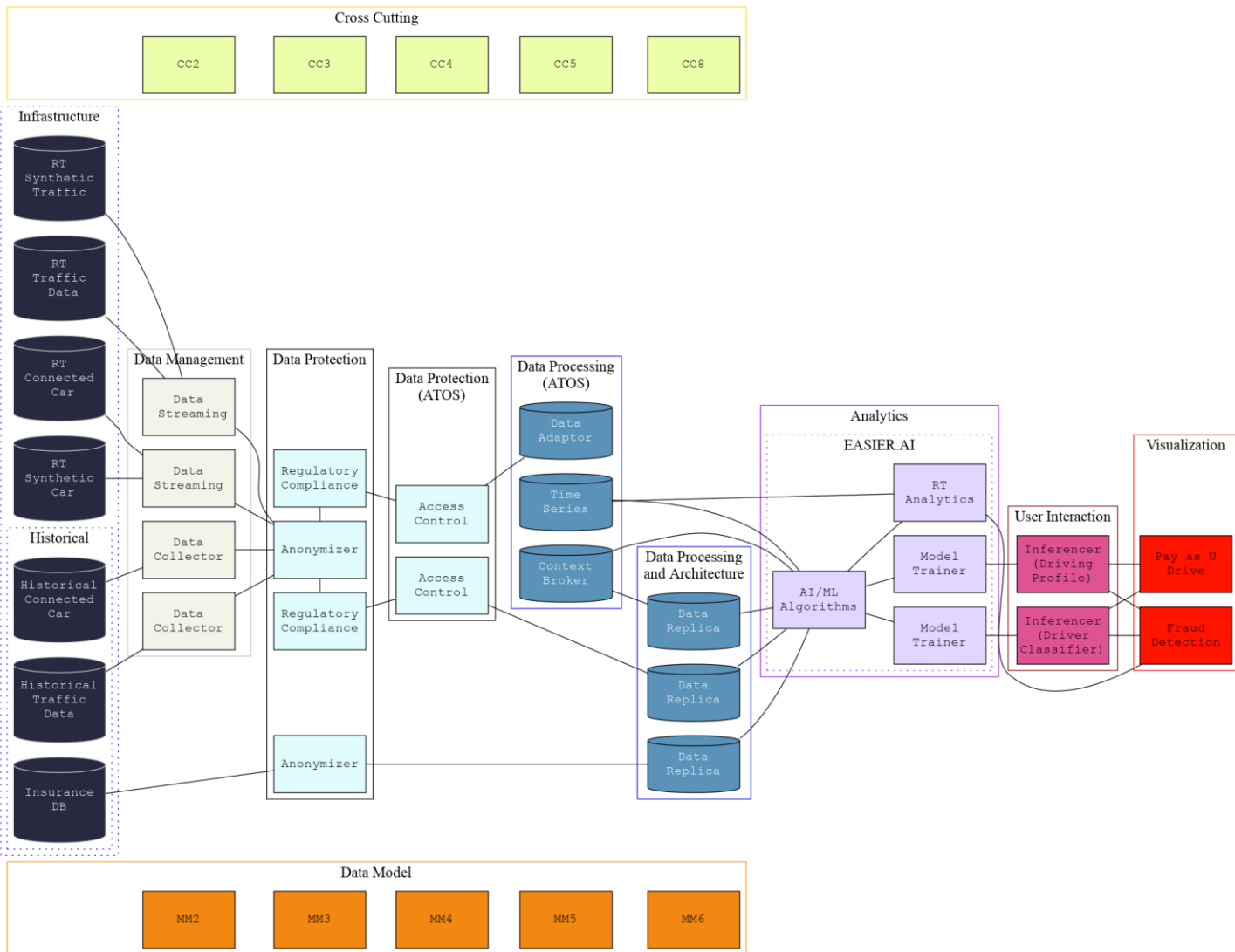


Figure 35: Pilot #11 Reference Architecture (From D2.13)

The pilot’s Reference Architecture (Figure 35) and main data flows have been presented and detailed in D2.13. This RA can be simplified considering:

- A Data Management layer, that selects, captures and curates the data sources required to implement the pilot’s functionalities. On this first stage, real connected vehicles and simulated traffic routes are the main implemented sources, assisted by weather information and traffic incidents collected for the area where the real vehicles will be driving.
- A Data Protection and Data Processing Layers in charge of homogenise and store all data collected, according specific data models (provided by FIWARE), so these are available for the analytics processes. Here are also included all the operations needed to anonymise/pseudoanonymise (as required) the captured data and protect this information from unauthorised accesses as well as data uploading from untrusted sources.
- An Analytics block, fed by the data layers, were different ML/DL technologies and visualization tools will enable data monitoring, analysis, and exploitation. Two main AI models (and inferencers) will be developed here, the Driving profiling and Driver Classifier tools (RA User Interaction), that will come up with the final services: “Pay as you Drive” and the “Fraud Detection” (in the RA’s Visualization layer).

The main stakeholders for this pilot are the insurance (car) companies and their insured drivers, who will exploit the driving profiles and drivers’ classifications and benefit from customised prices respectively. In Pilot #11 these stakeholders will be represented by Dynamis (DYN) that provides the user stories in D2.1 and D2.2.

To complete this pilot: Automotive Technology Centre of Galicia (CTAG) manages the real drivers' enrolment and real connected cars, plus traffic incidences around driving areas; Atos (ATOS) provides the pilot's core platform, including the traffic simulation tool and the weather datasets. It also develops the AI models to implement the final services; and Gradient (GRAD) that implements the Anonymization Tool and takes care of all the data managed within Pilot #11 to be GDPR compliant.

### 2.11.1 Technological components and Services

Going a step beyond the Pilot's RA towards the functional overview shown in Figure 34, the High Level Architecture presented in Figure 36 presents the software components that build the Pilot's use cases. This figure has been used in D6.1 to identify hardware requirements and in D2.5 to describe the technologies behind its principal components. This document links the shown software components with the corresponding RA layers, providing some details about their implementation. In this sense:

- **IoT infrastructures** supply raw datasets required for the implementation of the final services. The pilot has already identified and linked connected vehicles (real and simulated); weather stations (from AEMET); roads (from OpenStreetMap) and traffic alerts.
- **Data Collection & Aggregation** and **Data Normalization** components (Data Management and Protection layers of the RA): based on NGSI-LD and FIWARE Data models, define the rules to ingest data from IoT infrastructures. First functional versions for the identified IoT sources are deployed and ingesting data. Remark here the work done to integrate the Simulation of Urban Mobility (SUMO) tool with the Pilot's framework, following the NGSI guidelines. Also in this layer, Gradient's **Anonymizer** tool analyses and anonymises (when required) the collected data before being uploaded.
- **Connected Car framework** (Data Processing layer in RA): composed by the FIWARE Orion Context Broker<sup>11</sup>, that supports all context management functionalities (context information broker), and an instance of the FIWARE QuantumLeap General Enabler<sup>12</sup> (context information persistence) that supports historical information management. A first instance, covering these two components, has been implemented and deployed in Atos' infrastructure.
- **EASIER-AI** component (RA Analytics layer) is a Hybrid (Cloud/Edge) under development framework that facilitates to develop, measure, monitor and deploy customised AI models. It is built on top of the Elastic Search, Kibana and TensorFlow slate of three and enables different ML/DL technologies deployment. On top of this, Pilot #11 is developing (and will train) the Driving Profiling and Driver Classification inferencers (User Interaction RA layer) that will support the "**Pay as You Drive**" and the "**Fraud Detection**" services (Visualization RA Layer).
- The access to these frameworks (Connected Car and EASIER-AI) is protected by an OAuth **identification and authentication** component that relies on the FIWARE KeyRock IdM<sup>13</sup>. SSL/TLS is used to protect communications. This is deployed and integrated with the Connected Car framework.

---

<sup>11</sup> <https://fiware-orion.readthedocs.io/en/master/>

<sup>12</sup> <https://quantumleap.readthedocs.io/en/latest/>

<sup>13</sup> <https://fiware-idm.readthedocs.io/en/latest/>

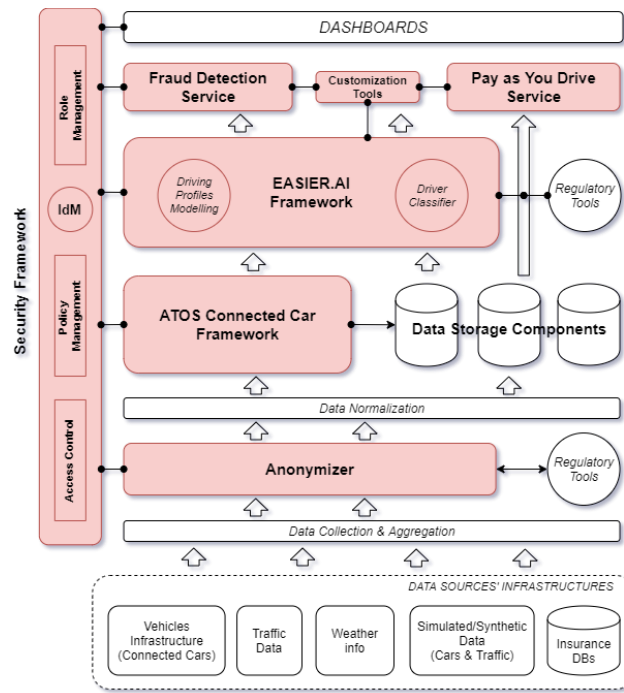


Figure 36: Pilot #11 High Level Architecture (updated from INFINITECH D6.1)

### 2.11.2 Data sets status

Within D2.5 in WP2 and D5.1 in WP5, the initial datasets to start Pilot #11 have been already documented and will be formally updated on their successive versions. This section presents an updated summary of the data that is put in place to start the pilot’s performance and AI initial analysis.

Collected information is homogenised and normalized (within Data Collection, Aggregation and Normalization layers) according predefined FIWARE data models, based on its NGSI protocol, before entering the Connected Car framework. The models currently used are:

- **Vehicle**<sup>14</sup> to map all (real and simulated) data captured from connected cars
- **WeatherObserved**<sup>15</sup> to capture data from AEMET’s weather stations (area of Vigo’s city)
- **Alert**<sup>16</sup> to get data from DGT’s reported traffic events (area of Vigo’s city)
- **Road**<sup>17</sup> and **RoadSegment**<sup>18</sup> that map information from roads and lanes (captured from OpenStreetMap<sup>19</sup>) were both, simulated and real connected cars are driving.

From D5.1 Pilot #11 table and aligned with these models, the pilot manages:

- **Current available datasets**
  - Connected Vehicles (from CTAG infrastructure): one On Board Unit (one connected vehicle device) is uploading to test CTAG Data Ingestion. (10 Mb)
  - Simulated Vehicles (from SUMO tool + Data Injector): one scenario reporting 30K vehicles (8 Gb)
  - Roads datasets (extracted from OpenStreetMap): 2 Gb for considered scenarios.

<sup>14</sup> <https://fiware-datamodels.readthedocs.io/en/latest/Transportation/Vehicle/Vehicle/doc/spec/index.html>

<sup>15</sup> <https://fiware-datamodels.readthedocs.io/en/latest/Weather/WeatherObserved/doc/spec/index.html>

<sup>16</sup> <https://fiware-datamodels.readthedocs.io/en/latest/Alert/doc/spec/index.html>

<sup>17</sup> <https://fiware-datamodels.readthedocs.io/en/latest/Transportation/Road/doc/spec/index.html>

<sup>18</sup> <https://fiware-datamodels.readthedocs.io/en/latest/Transportation/RoadSegment/doc/spec/index.html>

<sup>19</sup> [www.openstreetmap.org](http://www.openstreetmap.org)

- Weather info (from AEMET weather stations, captured by ATOS): 2k registries (1 Mb)
- Traffic Events (from DGT, captured by CTAG): 1k registries (0,5 Mb)
- **Under development:**
  - Historical datasets (CANBus data + NMEA) from real connected vehicles
  - Historical datasets from traffic events
  - Connected vehicles (from CTAG): CANBus + NMEA (location) from real connected vehicles (20 vehicles driving 4 hours a day for 1 year). To start in February 2021.
  - Car insurance data.

### 2.11.3 Testbed

Pilot’s #11 final deployment relies on UNINOVA infrastructure, as detailed in D6.1. Further details of the software/hardware first analysis and their results can be found in that document, but are summarised in 0.

Technical Components					
Component	Nodes	Cores by Node *	Memory (RAM) **	Storage ***	Special Hardware
EASIER.AI	1 Small + 2 Medium + 1 Large + 1 XXL	1x1 + 2x4 + 1x6 + 1x6 = 21 Cores	1x4 + 2x8 + 2x16 + 1x16 = 68 GB	1x100 + 2x500 + 1x2TB + 1x10TB = 13,1 TB	Some nodes would have access to GPU, FPGA, TPU, or other special AI hardware. AVX2 Instructions set support
Connected Car	1 Small + 2 Medium	1x1 + 2x4 = 9 Cores	1x4 + 2x8 = 20 GB	1x100 + 2x500 = 1,1 TB	
Anonymiser	1 Medium	1x4 = 4 Cores	1x16=16GB	1x1TB = 1TB	
Trainers/Inferencers AI	TBD later, in a first stage we will use the EASIER.AI infrastructure to also run trainers and inferencers				
<b>TOTAL</b>	9 Nodes	34 Cores	104 GB	15,2 TB	GPU, FPGA, TPU and AVX2 instructions support

\* Small Nodes = 1 Core; Medium Nodes = 4 Cores; Large Nodes = 6 Cores; XXL Node = 6 Cores

\*\* Small Node = 4GB; Medium Node = 8GB; Large Node = 16GB; XXL Node = 16GB

\*\*\* Small Node = 100GB; Medium Node = 500GB; Large Node = 2TB; XXL Node = 10TB

Table 14 Pilot’s #11 hardware testbed (first analysis). From D6.1

UNINOVA infrastructure it’s currently being dimensioned to provide support to several clusters, so it is not still available for deployments. Pilot #11 demonstrator is being deployed within ATOS premises. All first versions of the P#11 components follow an approach combining docker and kubernetes for their deployment to make easier the migration to the final testbed location.

### 2.11.4 Others non-technical requirements

Besides the technical requirements that compose the core pilot’s platform, the AI technologies deployment and the data collection, an additional and relevant requirement has been identified to obtain the best outcomes. This is related to the availability of enough data sources.

Anonymous Connected Cars vehicles, that will provide the routes (and vehicle data) needed to define, train and evolve the different AI models (and ML/DL technologies). The more vehicles enrolled, the better models obtained, but, on the same side, the more vehicles reporting around the same area, the better traffic models can be created and so, better driver classifications can be performed. In this line, the pilot will get 20 connected vehicles, mounting an smart on board unit that captures data from the CAN bus of the vehicle (technical vehicle data, such as speed, acceleration, systems status, etc.) plus an NMEA unit to capture GPS

vehicle’s location. These vehicles will start driving, supported by CTAG infrastructure, next Feb. 2021 and it is planned to report connected vehicles’ datasets for 4 hours a day and for at least 1 year long.

### 2.11.5 Implementation of a first Proof of Concept

First P#11 demonstrator is focused on data collection and homogenisation process, in order to identify any potential issue (or required data set) that may impact on the subsequent Pilot’s steps. This will also provide with fundamental elements the AI modelling stage. Following the initial approach on Figure 36 and architecture implementation on section 2.11.1, Figure 37 presents the functional diagram of the developed PoC.

As centred in data gathering, the ATOS Connected Car framework will be the core component to test and evolve. As mentioned above, this components’ set is mostly deployed in Atos Infrastructure, with support from CTAG to build and deploy their own data adaptors for their vehicles. In this sense:

- Data adaptors’ first versions (based on NGSI and FIWARE Data models) are deployed, ingesting data from the painted data sources.
- Connected Car core framework (Context Broker and Historical Repository based on FIWARE) is also ready, managing ingested context information. An NGSI-LD REST API is ready to access collected data.
- Identification and Authentication layer, based on FIWARE KeyRock IdM, is, in turn, managing Oauth tokens to grant access to the framework.

With all these components up & running, some dashboards are being developed in order to present the collected data and to start the data analytics processes. These will lead to identify the best AI approach to work on the Pilot’s final services.

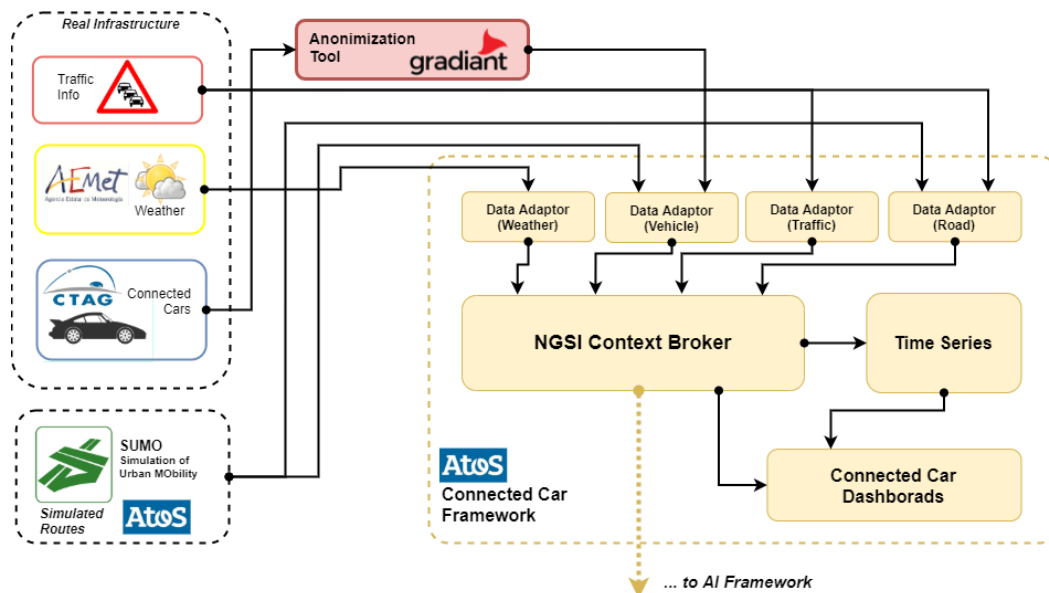


Figure 37: Pilot #11 Data collection PoC architecture

### 2.11.6 Next steps and timeline

Considering the PoC presented in section 2.11.5 as the current status, short-term next steps would include:

- Analysis and integration of the Anonymization Tool to grant GDPR compliance before ingest data from real connected vehicles (end of 2020)
- Ingestion of real connected vehicles (Feb. 2021).
- Analysis of simulated vehicle’s datasets to identify data correlations and first design of AI Models (end of 2020).
- Integration and deployment of the AI framework (Feb 2021).
- Implementation of AI models (March 2021).

### 2.11.7 Conclusions - Issues and Barriers

Based on the work done so far, the main foreseen challenges would include:

- Gathering enough and relevant datasets from vehicles that allow the system to define and detect a wide enough set of profiles that cover most of the driver’s population.
- Identify the proper correlations and relevant parameters from the collected datasets that better define and differentiate the profiles and so, the AI models to infer them.
- The mapping between the drivers’ profiles and the context information to provide accurate risks estimations.
- The availability of real datasets (connected cars) from insured drivers to match pilots’ services and exploit the results.

### 2.12 Pilot #12 Real world data for novel health insurance products

Pilot #12 focuses on health insurance and risk analysis by developing two AI-powered services: Risk assessment, that allows the insurance company to adapt prices by classifying the client according to their lifestyle; and the Fraud Detection which helps to identify fraudulent behaviour of the clients in using the activity trackers and answering the questionnaires. These two services rely on a people modelling that requires actual data and simulated persons to train. An overview of Pilot #12 is given in Figure 38:.

Current health insurance services are based on medical history and very static information. The innovation of Pilot #12 lies on applying new technologies (IoT and AI) to provide more dynamic and customized services.

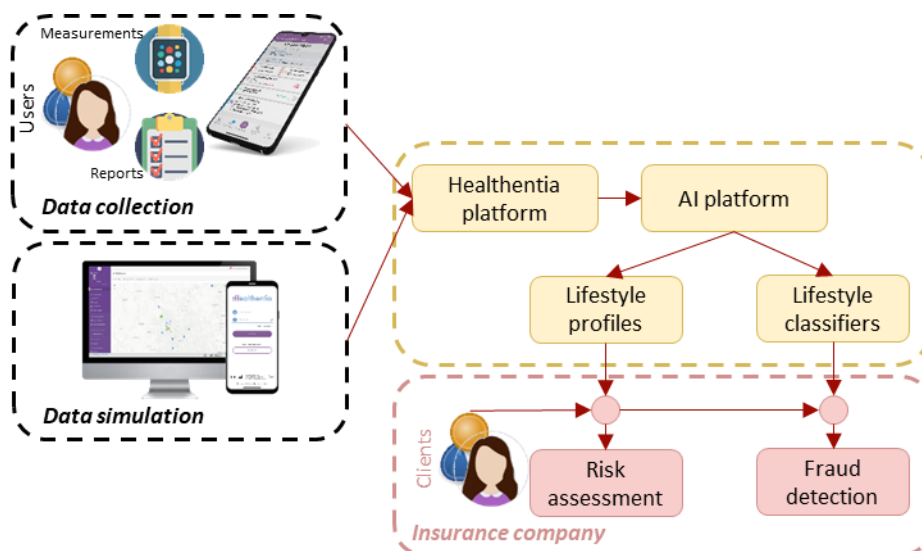


Figure 38: Pilot #12 Real world data for novel health insurance products overview

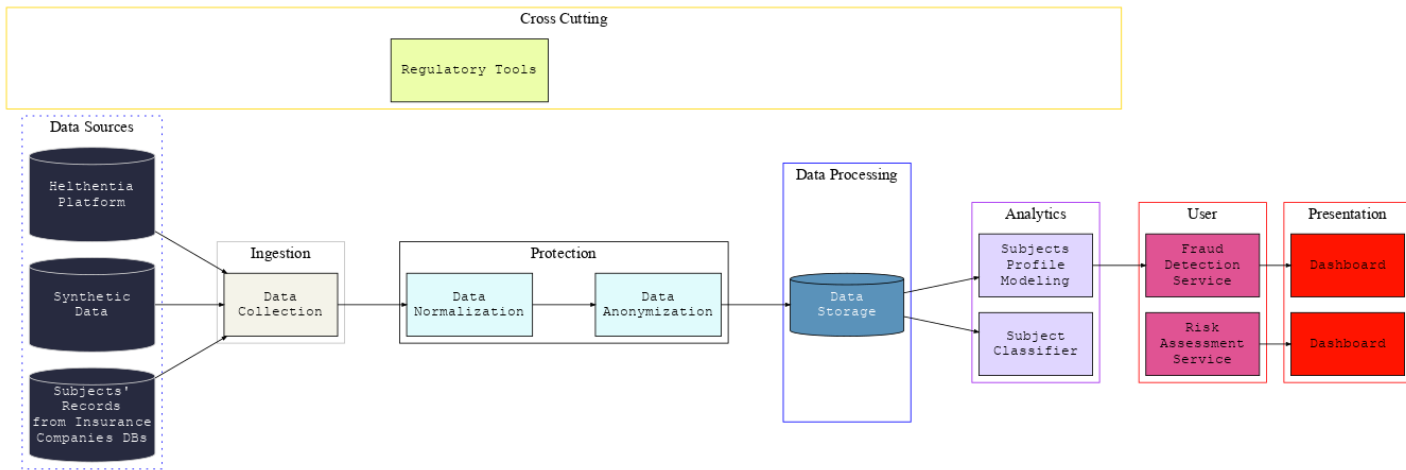


Figure 39: Pilot #12 Reference Architecture (From D2.13)

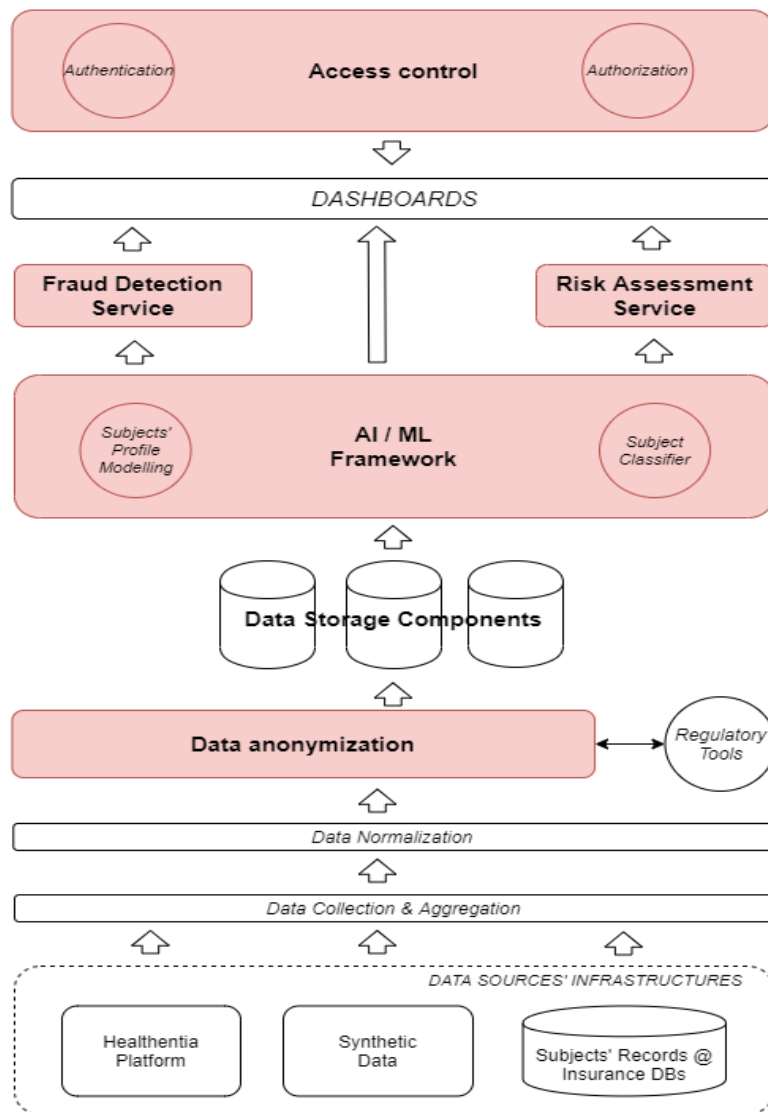


Figure 40: Pilot #12 High Level Architecture (updated From D6.1)

Pilot’s Reference Architecture (Figure 39:) and main data flows have been presented and detailed in D2.13. Aligned with Figure 38, this RA can be simplified considering:



- A Data Management layer, that selects, captures and curates the data from the actual and the simulated people.
- A Data Protection and Data Processing Layers in charge of homogenise and store all collected data, so these are available for the analytics processes. Here are also included all the operations needed to anonymise/pseudoanonymise (as required) the captured data and protect this information from unauthorised accesses.
- An Analytics layer, fed by the data layers, where different ML/DL technologies and visualization tools will enable data monitoring, analysis and exploitation. Two main AI models (and inferencers) will be developed here, the subject profiling and subject classifier tools, that will come up with the final services in the RA's Visualization layer.

The main stakeholders for this pilot are the health insurance companies and their insured clients, who will exploit the subjects' profiles and subjects' classifications and benefit from customised prices respectively. In Pilot #12 these stakeholders will be represented by Dynamis (DYN) that provides the user stories in D2.1 and D2.2. To complete this pilot: Roessingh Research and Development (RRD) manages the real subjects' enrolment; Innovation Sprint (iSprint) provides the data collection platform and the subject simulator. Singular Logic (SiLo) and Innovation Sprint (iSprint) develop the AI models to implement the final services; and Gradient (GRAD) implements the Anonymization Tool and takes care of all the data managed within Pilot #12 to be GDPR compliant.

### 2.12.1 Technological components and Services

The High-Level Architecture presented in Figure 40: presents the software components that build the Pilot's use cases. This figure has been used in D6.1 to identify hardware requirements and in D2.5 to describe the technologies behind its principal components. This document links the shown software components with the corresponding RA layers, providing some details about their implementation. In this sense:

- **IoT infrastructures** supply raw datasets required for the implementation of the final services. The pilot has already identified and linked the Healthentia platform for Real-World Data collection and the RWD Simulator, both from iSprint as the data sources. Clients' records maintained by the insurance companies are still under investigation.
- **Data Collection & Aggregation** and **Data Normalization** components (Data Management and Protection layers of the RA): Healthentia platform already handles ingestion from IoT infrastructures. Remark here the work done to implement and integrate the Real World Data (RWD) Simulator. Also, Gradient's **Anonymizer** tool analyses and anonymises (when required) the collected data before being uploaded.
- The **ML** component (RA Analytics layer) builds upon Scikit-Learn and Keras/TensorFlow the subjects' profiling and subjects' classification inferencers (User Interaction RA layer) that will support the two services (Visualization RA Layer).

### 2.12.2 Data sets status

Within D2.5 in WP2 and D5.13 in WP5, the initial datasets to start Pilot #12 have been already documented and will be formally updated on their successive versions. This section presents an updated summary of the data that is put in place to start the pilot's performance and AI initial analysis:

- An evolving dataset from Pilot 12 actual users. Currently there are 19 users providing measured and reported data. They all have enrolled and use diverse sensing systems (Fitbit or Garmin activity trackers, Apple Health Kit or proprietary Android sensing system). Not all uses actively contribute physical data and questionnaire answers, hence the amount of data each user has been contributing varies. Currently

9 of the users make sure their physical data are being reported daily, while 486 questionnaires have been answered, but 332 of them by two of the users. An example of user data from the PoC is given in Figure 41: .

- A dataset from simulated users. The simulator is not yet complete at the time of authoring this deliverable, but it is quite mature. It produces activity data in the form of daily aggregations and detected activities (similar to those collected from actual users with Fitbit activity trackers, Apple Health Kit and Android sensing), as well as intra-day ‘measurements’ (similar to those collected from actual users with Garmin activity trackers). It also produces answers regarding symptoms, liquids and meals, as well as answers to the weekly questionnaires on Quality of Life. Simulating 300 users for 3 months and 2 years (the 3-month period is for the activity trends to settle, while the remaining 2 years facilitate model building and service provision) currently results in 6.6GB of data.

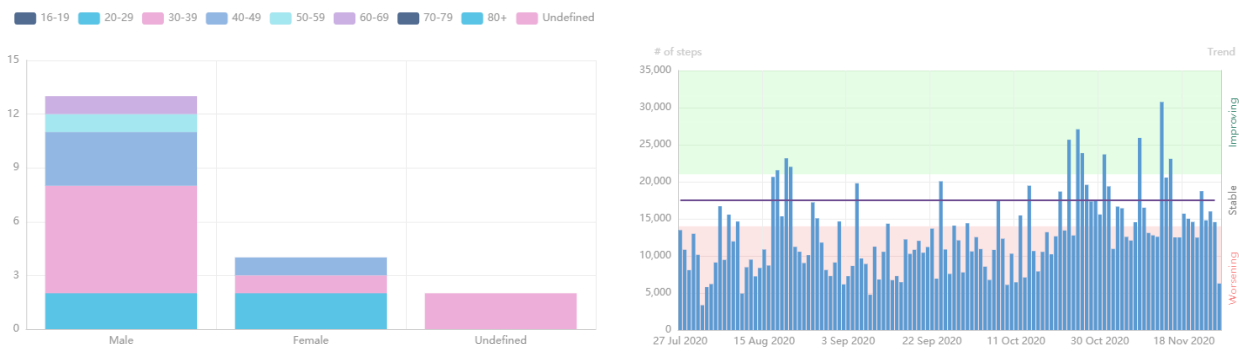


Figure 41: Pilot #12 PoC user participation. Users’ demographics (left) and a particular user’s steps data (right).

### 2.12.3 Testbed

Pilot’s #12 deployment relies on UNINOVA infrastructure, as detailed in D6.1. Further details of the software/hardware first analysis and their results can be found in that document, but are summarised in the following table:

Platform	Processors	Cores	Memory	Storage	Special Hardware
ML server	1	6 Cores / processor	32 GB	250GB	Access to GPU special AI hardware is very welcome
Application server	1	6 Cores / processor	16 GB	250GB	None
Anonymization server	1	4 Cores / processor	16 GB	500GB	None
<b>TOTAL</b>	<b>4</b>	<b>18</b>	<b>64 GB</b>	<b>1TB</b>	

Table 15 Pilot’s #12 hardware testbed

UNINOVA infrastructure it’s currently being dimensioned to provide support to several clusters, so it is not still available for deployments. Pilot’s #12 demonstrator is currently being designed following a docker+kubernetes approach for their deployment to facilitate possible initial deployment at a temporary server and final deployment at the UNINOVA testbed.

## 2.12.4 Other non-technical requirements

The success of pilot #12 depends on the wealth of data made available for training and inference. Data is obtained from users employing measurement devices and the will to participate in the pilot by using the measurement infrastructure, reporting symptoms, liquids and meals, and answering the questionnaires. The RWD Simulator is being built to fill in the necessary data volume, but its models depend on the observations made on the actual data being collected.

## 2.12.5 Implementation of a first Proof of Concept

The primary focus of the Proof of Concept demonstrator of Pilot #12 is data collection: what to measure, what to ask for, how to collect and how to simulate. Secondary points of focus are the pilot's testbed and the risk analysis service.

Pilot 12 data collection is based on the Healthentia platform, by Innovation Sprint. Healthentia is an eClinical system that comprises mobile apps at the data source (the pilot participants and a platform for collecting the data. A postal app allows data visualization. The pilot's first goal has been to repurpose Healthentia from the clinical to the health insurance domain. To this extend, the data collected, and the questionnaires forwarded to the pilot participants have been selected and defined. Currently we collect physiological data from four possible sources, a Garmin connector, a Fitbit connector, an Apple Health Kit connector and a proprietary Android sensing service. Our questionnaires span symptoms, liquid and food consumption and the self-assessment of quality of life and health, the EQ-5D-5L questionnaire.

Data are also being provided by the RWD Simulator built for INFINITECH. The simulator accepts people's personality traits and health profiles whilst simulates their activities and the corresponding measurements and questionnaire answers. The simulator data have the exact same structure as the actual ones and are also collected by Healthentia.

Regarding the pilot's testbed, a temporary setup is being managed by Innovation Sprint using a Linux 2020LTS server at Hetzner. It is a VM with 2 vCPUs, 8 GB RAM and 80 GB storage (CX31 instance). There Ubitech's Data Capturing Tool has been configured to capture data from the Healthentia API and store it in the LeanXcale DB.

Finally, regarding the risk analysis service, classifiers have been built using the simulated data to predict if the health of a person is expected to improve or not during a week, based on the week's measurements and reports. Both Random Forest and fully connected Neural Networks classifiers have been trained, with the NN one performing slightly better, achieving 78% correct identification of the health trend.

## 2.12.6 Next steps and timeline

As presented in D10.1, Pilot #12 consists of four phases:

- Data Collection Proof of Concept Phase (M11-M13), described in the previous section.
- Continuous Data Collection (M11-M36), running throughout the duration of the pilot, collecting data to train models.
- Data Sharing Acceptance and Usability Study (M14-M18), addressing user aspects.
- Validation and Evaluation (M24-M36), running the pilot with actual users and offering the services to the end user insurance companies.

The medium-term steps (Q4 2020) involve:

- Utilise the stored data in LeanXcale to get predictions from the current health trend classifier.

- Finalise and fully integrate the RWD simulator.
- Add to the testbed the Anonymization Tool by GRADIANT

The longer term (Q1 2021) implementations involve:

- Port the system to UNINOVA testbed.
- Run the Data Sharing Acceptance and Usability Study and utilise its results,
- Based on the current health trend classifiers, work towards the risk assessment service,
- Derive the fraud detection service.
- Provide the analytics platform to aid end-user decision making.

### 2.12.7 Conclusions - Issues and Barriers

The PoC of Pilot 12 allowed us to implement the data collection system, addressing both, the what and the how. The low engagement of the PoC participants (about 50%) is alarming and will be addressed in the Data Sharing Acceptance and Usability Study from the data, privacy and UI/UX aspects. It is also being addressed technically by increasing the measurement options and optimising the Android sensing service. Our aim is to be gathering soon enough and relevant data from diverse users to facilitate both risk assessment and fraud detection services.

The testbed will be transferred to its permanent location at the NOVA server, but the PoC already set in motion all the collaborations necessary for its setup amongst the INFINITECH partners not members of the pilot.

The risk assessment service has been addressed at the PoC via an initial predictor of weekly variations of health. Both classifiers and regressors will be built in the coming months, the feature vector used to train them will be optimised and as a result the heart of the service will be in place. The fraud detection has not been addressed yet, and this is a concern, since today people cheat on their activity trackers just to get a badge in their favourite wellness app. This could escalate when health insurance discounts are involved.

## 2.13 Pilot #13 Alternative and automated insurance risk selection and insurance product recommendation for SME's

Pilot #13 will monitor risk's changes, so it will be able to radically improve the risk management that companies (SMEs) face in the development of their daily activity. The indicators will be based on information from each of the companies coming from online sources that will give information about the digital presence and activity of those companies like activity, business volume, participation in social networks, number of employees, use of ecommerce, payment platform etc, etc. The company to be analysed does not need to provide many information, developed tools are in charge of searching and gathering information related to his company from many sources. In this way, risk profiles of each of the companies analysed will be generated, allowing to customize the product offering and to make a permanent automated risk management. But this is not the only usage of data, insurance companies will use these information, resulting on better customized products.

An overview of Pilot #13 is given in the Figure 42: and Figure 43:

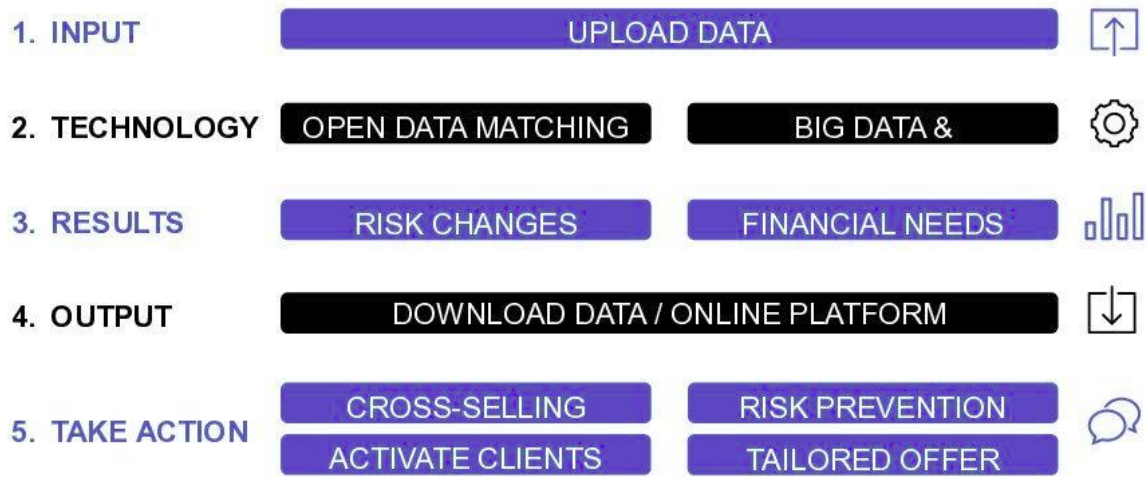


Figure 42: Pilot #13 summary

The following figure provides an INFINITECH-RA compliant logical view of the logical architecture of the pilot, from D2.13.

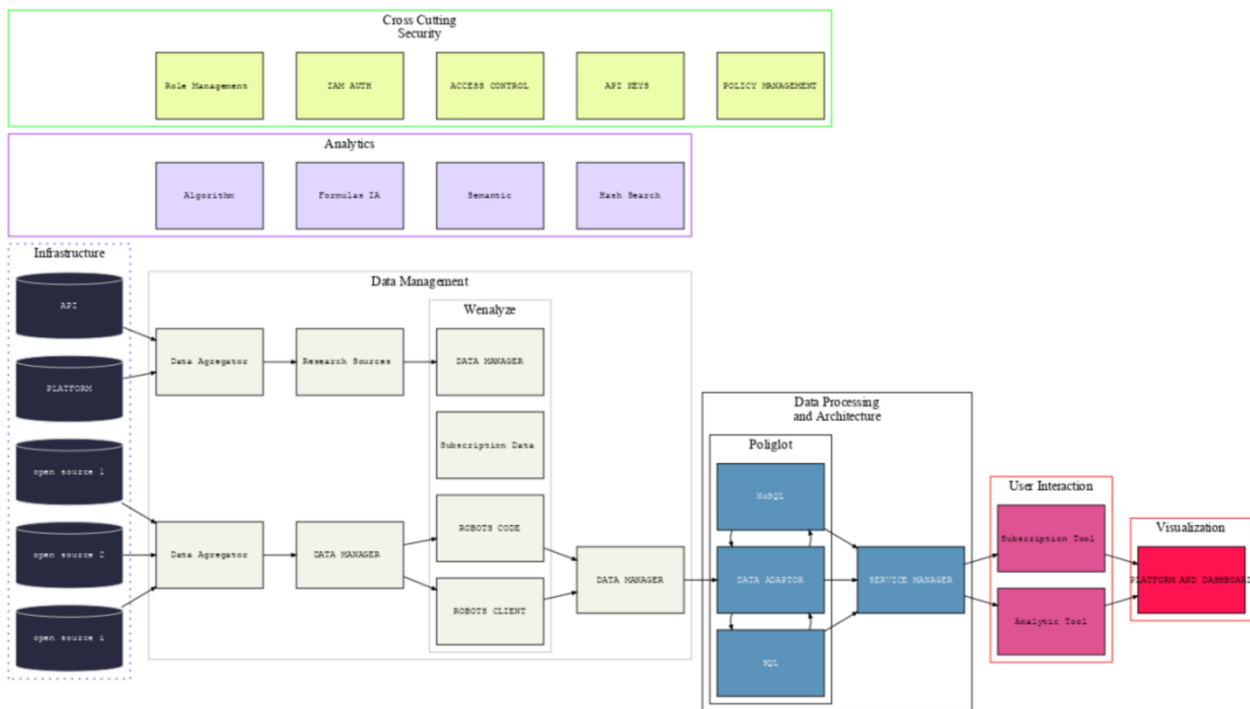


Figure 43: Pilot #13 Reference Architecture (from D2.13)

Pilot’s Reference Architecture and main data flows have been presented and detailed in D2.13. In summary, the main components to be developed in this pilot are:

- Data Sources layer, that through information collection, select and obtain the information from dozens of sources in an efficient way, minimizing the necessary computation.
- A Data Management layer, that selects, captures and curates the data sources required to implement the pilot’s functionalities. This information management allows data collection sources to efficiently dump the data into non-relational streaming databases
- An Analytics block, fed by the data layers, where different ML/DL technologies and visualization tools will enable data monitoring, analysis and exploitation. Two main AI models will be developed to cover pilot’s uses cases

In pilot 13 the main participants are Wenalyze and LeanXcale. LeanXcale will provide the Data Management and Data Processing components, while Wenalyze will provide the infrastructure for obtaining data from open sources, the Analytics part, User interaction and the Visualization part.

The end users of the information obtained and processed by the platform are insurance and reinsurance companies and banks. Banks regard the sale, underwriting and control of risks from their business clients and SMEs. First contacts with different insurance companies have already been made. The comments have been very positive, and the pilots would start once the algorithms and the platform are implemented.

Also, a communication and conversion funnel has been created and is being distributed in both European Union countries and the United States. At present, 16 insurance and banking entities from eight countries are in this funnel and are in the process of marketing qualified lead to sales qualified lead. Different proofs of concept have been already agreed. The actual conversion funnel regarding end user is (Figure 44: ):



## Customer Acquisition Funnel

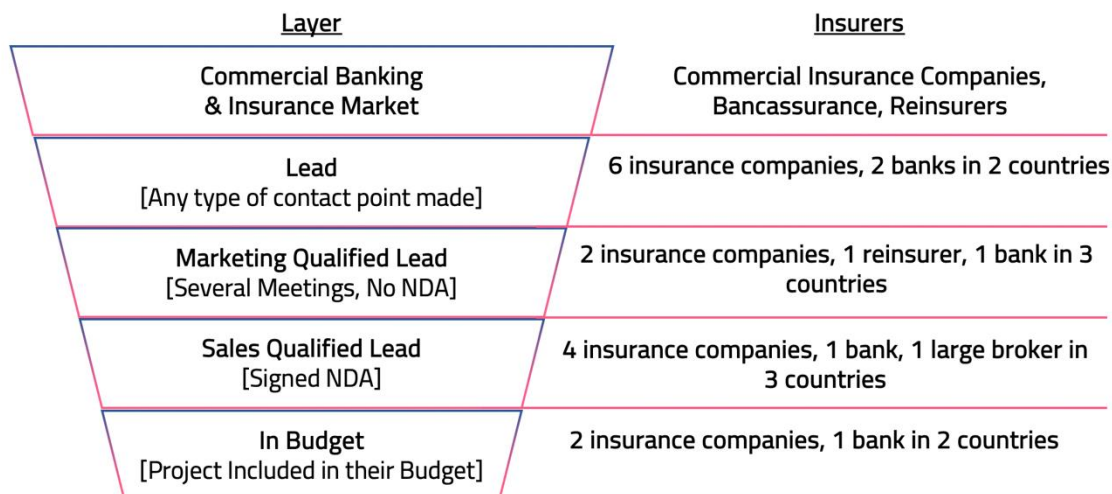


Figure 44: Pilot #13 Customer Acquisition Funnel

The development of the pilot is very positive, and it is expected to be completed in the time foreseen by the consortium

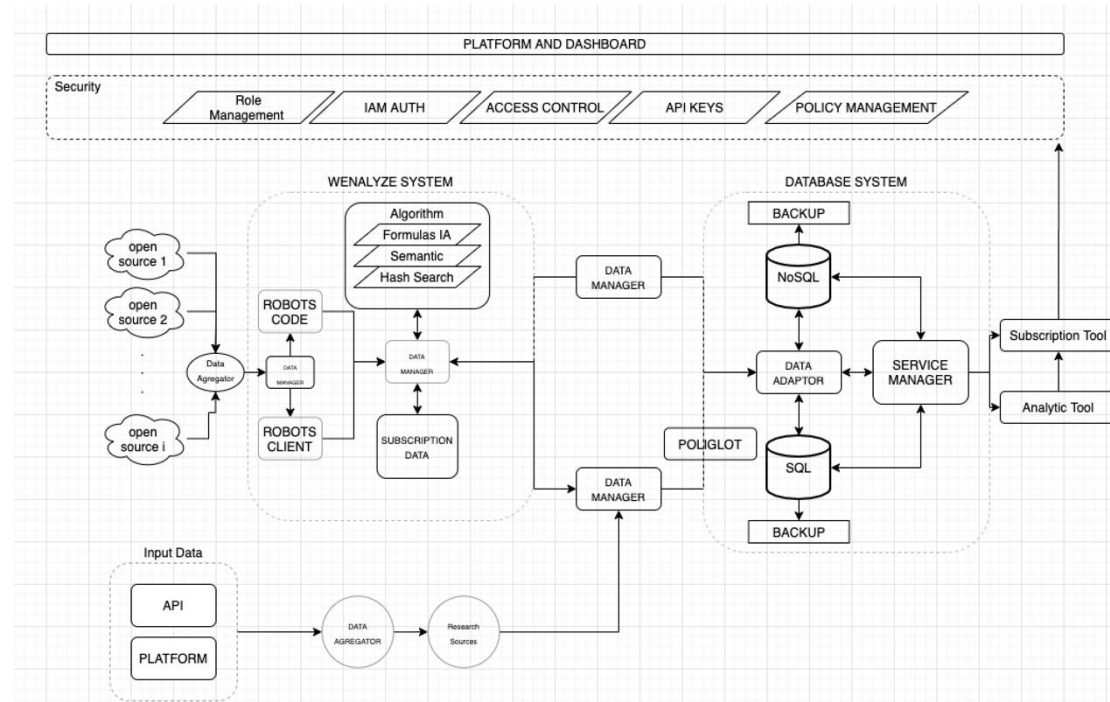


Figure 45: Pilot #13 Platform schema

### 2.13.1 Technological components and Services

This document links the shown software components with the corresponding RA layers, providing some details about their implementation (Figure 45: )

- Data Sources (infrastructure). To obtain the data from the information sources we will use the automatons developed by Wenalyze, based on extensions and instances
- Data Management (Data Collection and aggregation), For data management we will rely on LeanXcale with its non-relational databases, its data manager and its polyglot
- Analytics, For the application of the models developed in ML we will use the Wenalyze platform that will connect with the testbed in NOVA through Javascript
- Connectivity, Finally the connectivity is foreseen through an API rest, but to facilitate the realization of the PoC also has developed connectivity through the use of browser, access by users and password or the upload of files in CSV format

### 2.13.2 Data sets status

Within D2.5 in WP2 and D5.13 in WP5, the initial datasets to start Pilot #13 have been already documented and will be formally updated on their successive versions. This section presents an updated summary of the data that is put in place to start the pilot’s performance and ML initial analysis.

Dataset Name	Dataset (short) description	License/ Privacy	Anonymized	Data Type
SMEWIF	SMEs website information and functionalities. Description of the text containing in the website of the companies, services and structure of the company	Open	No	Text
ROPS	Review and opinions platforms. Reputation information and opinions of clientes about productos ans services	Open	No	Text
EUBD	European SMEs Business Directories. Oficial and legal information about the comanies, social object, activitie, other companies where they have equity	Open	No	Text
GIO	SMEs geolocation information and characteristics, images and geographical information	Open	No	Text / Image
SMSIP	Social media SMEs information and presence.	Open	No	Text

Table 16 Pilot #13 Data Sets Details

The datasets are not yet available on the testbed, but it is planned that they will be available on the M14 and from then on, they will gather the information available in the sources for the elaboration of the models

### 2.13.3 Testbed

The Pilot 13 will implement an automation of the subscription process that helps the insurance company reduce costs. In addition, being able to verify that the data entered is correct with a double verification avoids possible errors in the cost of the insurance premium The monitoring and identification of real-time risk

changes allows the company to know if the insurance cost really corresponds to the real risk of the SME or if it should increase or decrease it to adapt it to its current situation. The infrastructure that will be used will be place in UNINOVA, as detailed in D6.1

- In Nova hosting just will be implement the data base system by LeanXcale
- The solution consists in make transactions and be storage and manage in non-relational databases

Hw Requirements	CPU's 8 cores	Architecture, X86_64	Memory, 32 Gb of RAM	Storage, 500 Gb	Storage type, SSD	Network performance 25 Gigabit
Sw Requirements	Linux operating system (Debian like Ubuntu, no GNU)	SSH access	Node.js and NPM	GITHub	NginX server	

Table 17 Pilot #13 Hardware/Software Requirements

### 2.13.4 Other non-technical requirements

A wealth and variety of important data is essential for the proper functioning of pilot #13. The data are obtained from open sources through which we can obtain different information related to the real time activity of the companies. We will use two sources of data, one that is available internationally and a second that must be incorporated in each of the countries. These secondary data sources per country are not always available with the same information and this, although avoidable, can complicate the development of the pilot.

### 2.13.5 Implementation of a first Proof of Concept

This service makes it possible to monitor the risks of SMEs now and in the future and therefore improves the control of the accident rate, the renewal of insurance policies and offers personalised insurance cover

The RA that will be used are Data processing and data analytics. Related to the information that will be use as a input in the proof of concept will be, SMEs website data, opinion platforms, business directories, social media and ecommerce platforms. The PoC will be run based on data recovered from Spanish companies



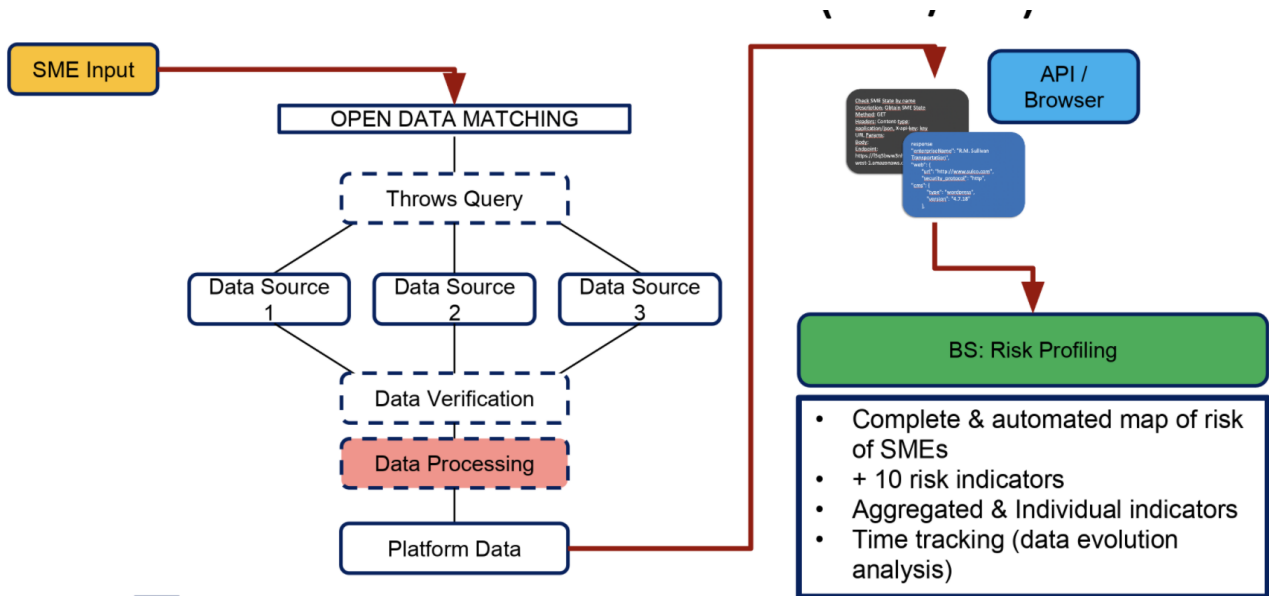


Figure 46: Pilot #13 Implemented Proof of Concept

### 2.13.6 Next steps and timeline

Based on what is reflected in the deliverables, pilot #13 develops its content in the following phases

- Data Collection & proof of Concept Phase (M10-M13) described in the previous section.
- Continuous Data Collection (M11-M36), running throughout the duration of the pilot, collecting data to train models.
- Data modelling and Usability Study (M14-M18), addressing user aspects.
- Establishing connectivity with insurance companies (API / CSV / Web browsing) (M15-M16)
- Contacts with possible end users to test the solution ((M18-M19) based on the funnel conversion designed
- Validation and Evaluation (M20-M25), running the pilot with actual users and offering the services to the end user insurance companies.

### 2.13.7 Conclusions - Issues and Barriers

At this moment, the Pilot #13 is progressing according to the plan. The intake of information is neither presenting any problem, nor the construction of the models.

The only delay with respect to the plan is the transfer of the AWS development to the Nova testbed. For the time being, this is being managed by LeanXcale, this delay should not impact on the deadlines set for the pilot.

Finally, to point out the efforts on the commercial promotion of the tool, starting with communication in forums of the sector in the European Union and obtaining the first leads for the conversion funnel.

## 2.14 Pilot #14 Big Data and IoT for the Agricultural Insurance Industry

The objective of Pilot #14 “Big Data and IoT for the Agricultural Insurance Industry” is to deliver a commercial service module that will enable insurance companies to exploit the untapped market potential of Agricultural Insurance (AgI), taking advantage of innovations in Earth Observation (EO), weather intelligence & ICT technology. EO will be used to develop the data products that will act as a complementary source to the information used by insurance companies to design their products and assess the risk of natural disasters. Weather intelligence based on data assimilation, numerical weather prediction and ensemble seasonal forecasting will be used to verify the occurrence of catastrophic weather events and to predict future perils that could threaten the portfolio of an agricultural insurance company. The INFINITECH AgI-module derived indices will allow and enable the agricultural insurance industry to enlarge its market, while delivering a larger portfolio of products at lower costs and serve areas, where classical insurance products could not be delivered. This is the main business and market innovation of this pilot.

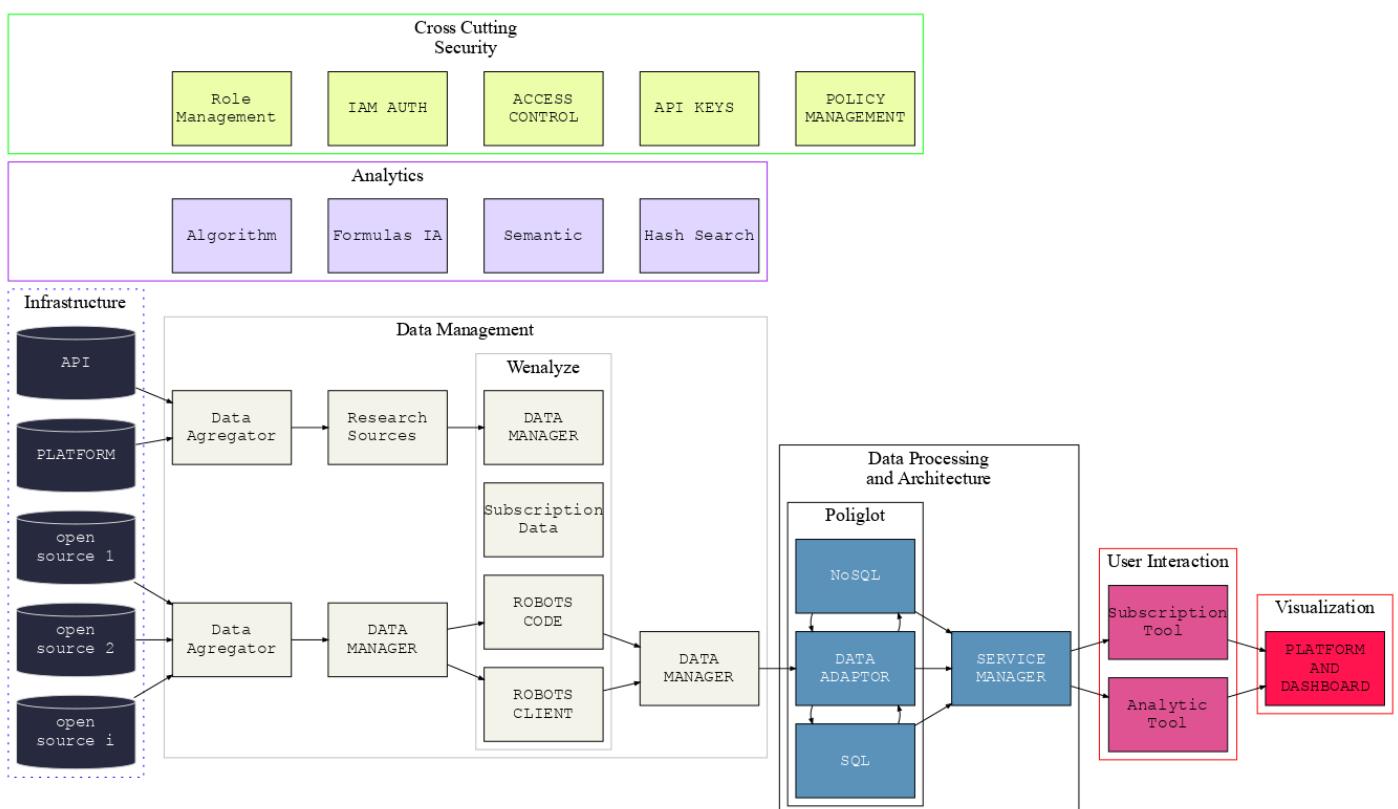


Figure 47: Pilot #14 Reference Architecture

AgroApps is developing the entire infrastructure for the pilot #14 data products, based on the reference architecture (Figure 47:) starting from data collection from different sources over processing and analytics to user interface & data visualization. The ongoing development of the service module is based on scientific research in the field of agricultural insurance, climate & weather risk modelling and the most recent evolutions in the area of remote sensing technologies. The reason for this is that these three areas will play a crucial role for the future of agricultural insurance providers in order to tap new markets, provide better risk transfer solutions and make insight-based strategic decisions. To meet the demands of this rapidly evolving field, it is necessary to follow these current developments.

As described in the User Stories (D2.1), the service module is mainly designed for staff working in the underwriting and sales department of agricultural insurance companies (majority of User stories serves this group of end-users). However, within this departments, there are several roles who can benefit from the

services provided by Pilot #14. First of all, Actuaries (business professional/mathematician who analyzes the financial consequences of risk by using statistics) are able to improve their data set for risk pricing and product development based on the data retrieved from the service module. Based on this information, Underwriters can better evaluate the risk and exposure of potential clients (crop monitoring) and hence make the overall insurance portfolio more resilient by at the same time increasing the outreach to clients (farmers). Additionally, Sales Agents can identify areas where to prioritize sales activities without increasing the cumulative risk since they are aware of e.g. regional risk profiles.

Lastly, with the support of data derived from the Octopush EO (damage assessment services), loss adjusters have additional information to make the on-farm process of loss adjusting more efficient and for certain perils conduct this process remotely via the service module (without visiting the farm/respective field).

In addition to the implementation within insurance companies, at a later stage of the project other users in the insurance value chain can also be considered as end users (see INFINITECH deliverable D2.1).

A first contact inside an insurance company in the Area of Interest (Serbia) has been made and immediately generated interest because of the benefits the Pilot #14 service module has to offer. The feedback on the presented services was very positive, just a final decision by the management is outstanding.

As in this very first stage of the preparation of the pilot site the receiving of an appropriate and high-quality dataset from the pilot insurance company and the application of the services described in Pilot #14 have highest priority, there are no training plans developed for deployment to the final user yet.

However, for the internal deployment at the final pilot site, Pilot #14 can provide an independent web-based user interface for the end users to access the service module via their browser.

### 2.14.1 Technological components and Services

Based on the reference architecture (D2.13, Figure 47) the following components and services will be deployed and used as part of the pilot:

#### ICT Modules

- Octopush EO Service: Octopush EO Service is an integrated satellite derived software service, which collects earth observation, geospatial, in-situ and other geo-referenced data, it applies appropriate processing algorithms and returns the results in a ready-to-use format.
- AgroApps Weather Intelligence Engine (AgroApps WIE): The WIE is an integrated weather derived software service which collects weather information from several resources and along with the geo-referenced data, it applies appropriate processing algorithms and returns the results in a ready-to-use format.
- Data integrator: The Data Integrator acts as a bridge between the WebGIS subsystem, Octopush EO service and WIE. It is responsible for performing the essential scheduled calls to the data providers in order to fetch and process the desired EO and weather information. It is able to run calls on demand or daily data integration tasks by retrieving EO data and weather products from Octopush EO service and WIE and transforms, binds, injects those into the WebGIS database.
- Business and Geospatial DB: Business DB offers a storage layer essential to carry the business logic and relevant information/ data stored and managed by API. It also stores, retrieves and provides information related to user accounts, settings, actions and preferences. The geospatial data storage and data

persistence mechanisms allows the storage of the geometries and zonal statistics and provides the essential functionality for querying and retrieving data via an API or WMP server components.

- Web Map Server (WMS Server): WMS is responsible for rendering and serving of the GIS layers to the User Interface.
- RESTful API: The API will act as a communication and data exchange bridge, that allows the platform to share processed and structured content internally, between the different components.
- User interface: The front-end user interface is the gateway responsible to present all the system data through user-friendly controls and web mapping interfaces.

### Services for the Insurance Sector

- Remote Damage Assessment for drought and hail (TRL 6):
- Flood and wildfires mapping (TRL 6):
- Short and medium range weather forecasts (TRL 9):
- Seasonal Climate Forecasts of Agroclimatic Indicators (TRL 8):
- Climate Risk Assessment (TRL 9):

### 2.14.2 Data sets status

The main data source for the pilot is produced by satellites and the weather intelligence engine. The Earth Observation (EO) data will be derived from the satellites Sentinel-1,2,3, LandSat-8, MODIS and PROBA-V. Also, numerical weather predictions for the pilot areas (gridded data) are generated each day and will replace the previous prediction. Lastly, gridded climate indices based on ERA-5 Land and ERA-5 Reanalysis Data will be used for the pilot.

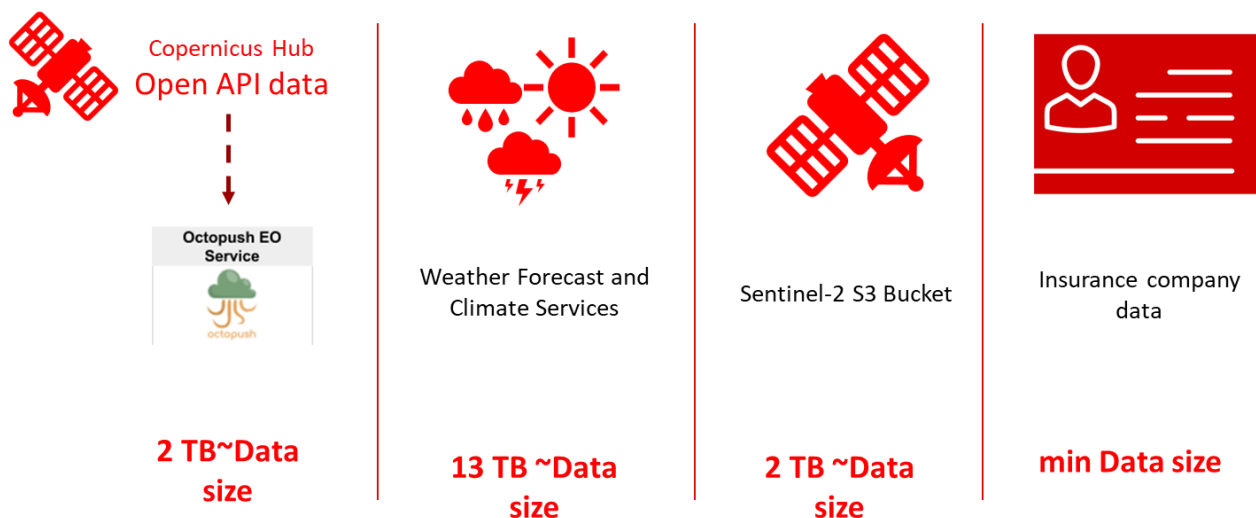


Figure 48: Pilot #14 Data sources

The data produced will result in a solution for Agricultural Insurance companies allowing them to efficiently couple EO satellite data and weather/climate data with any type of complementary data (from separated drone shots to ultra-high-resolution SAR imagery). The INFINITECH AgI module will enable Insurance companies to alleviate the effect of weather uncertainty when estimating risk for AgI products, reduce the number of on-site visits for claim verification, reduce operational & administrative costs for monitoring of insured indexes and contract handling, & design more accurate & personalized contracts. By deriving

impartial indices on top of a multitude of data, the module will allow insurers to reduce significantly the time needed for handling and verification of claims and the costs imposed by fraud, moral hazard and adverse selection. In D5.1 (WP5), the available datasets of Pilot 14 have been already documented.

### 2.14.3 Testbed

All modules of Pilot 14 services will be hosted in AgroApps premises, except the weather intelligence engine that will be deployed in UNINOVA's infrastructure. In this sense, server specifications will be defined at a later stage.

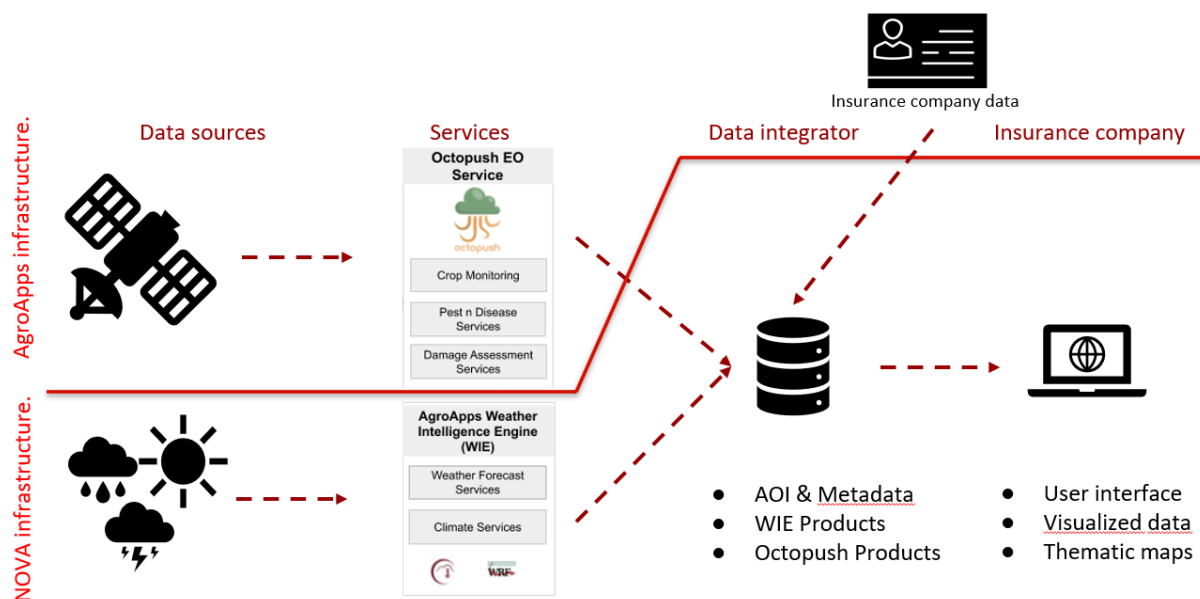


Figure 49: Pilot #14 testbed

### 2.14.4 Others non-technical requirements

Besides the technical requirements for the pilot, there are also other non-technical requirements in order to test the application successfully. These requirements mainly relate to the data provided by the agricultural insurance company:

- In the past, we observed that the quality of the data provided by agricultural insurers was often poor. This is mainly due to the IT structure of the insurers, which often does not allow targeted queries at short notice.
- However, in order to apply the structured and unstructured data provided in the shared testbed by AgroApps and UNINOVA (Earth Observation (EO), Numerical Weather Prediction Data, Reanalysis Data and Seasonal Climate Forecasts) to the insured/to be insured regions, the data provided by the insurance company needs to be accurate, timely and on a correlating spatial resolution.
- This applies not only for the clear identification of a region/field by coordinates or IDs from national databases, but also to the existing/desired form of insurance cover, the crop to be insured, average yield values and a (potential) loss history.
- If the data quality is insufficient, the national statistics office could also be consulted, e.g. for average yield data.

Furthermore, Pilot #14 make use of the respective national network of Weather stations for collecting data used in the Weather Intelligence Engine to predict weather and climate patterns.

### 2.14.5 Implementation of a first Proof of Concept

The first Pilot #14 Proof of Concept (PoC) will focus on a data processing architecture and a data analytics infrastructure to create an Area Risk Profile for the defined Area of Interest (AOI) in order to assess the risk of natural disasters and to develop a pricing framework for a drought index product (Figure 50)

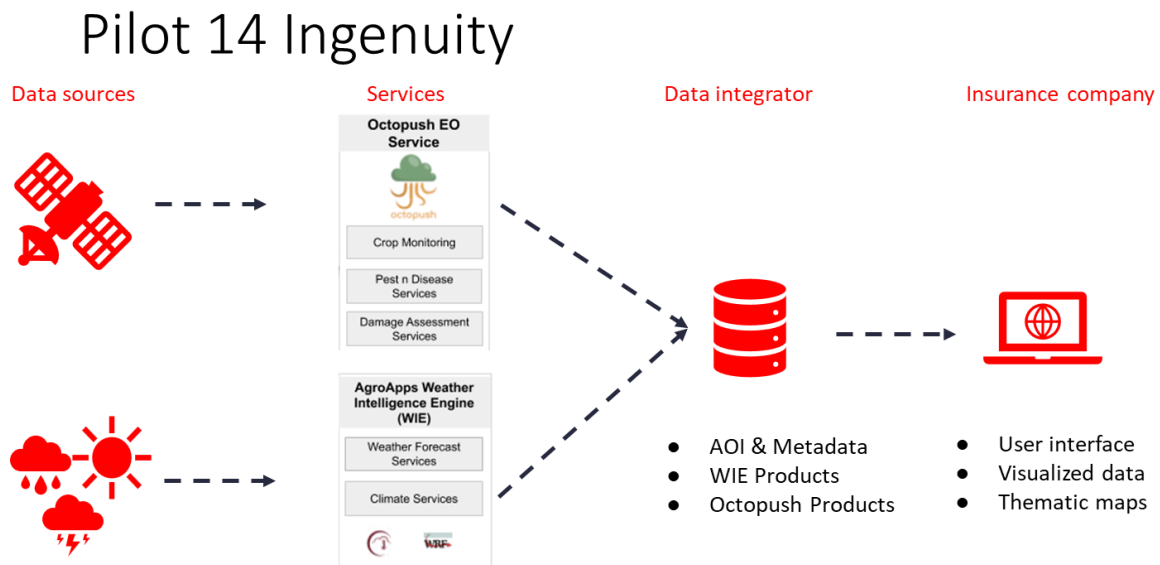


Figure 50: Pilot #14 Proof of Concept

Therefore, EO data derived from satellites and weather intelligence based on data assimilation, numerical weather prediction and ensemble seasonal forecasting will be used to verify the occurrence of catastrophic weather events and to predict future perils that could threaten the portfolio of an agricultural insurance company (Figure 51).

In this first phase of the pilot site preparation, the focus will be on providing solutions for users situated in agricultural insurance companies (Actuaries, Underwriters, Sales Agents, Loss Adjuster) as described in the User Stories #14.01-14.08 in deliverable D2.1. and above).

By combining the components (Figure 50) developed in the AgroApps and UNINOVA Infrastructure and the data set from the insurance company (2.14.4), the respective user application can be set up and tested. The results of this first PoC will help to improve the data flow and data analytics processes for the Pilot’s final services.

Term	Index
Insured value	<b>Index value</b> , derived from historical average (high correlation with field variable(s) (e.g. <b>weather or yield data</b> ))
Client risk retention	% of historical average (area above trigger)
Payout limit	Maximum <b>payout</b> : % of historical average (area between trigger and exit)
Framework for index insurance pricing	

Figure 51: Pilot #14 Data requirements for developing a pricing framework

### 2.14.6 Next steps and timeline

*Next development steps and timeline to implement all the user stories and components*

After setting up the monitoring (Indices) and data collection including in-situ data (Octopush API) for the AOI in August and September, further data sources (Weather Intelligence Engine) were integrated by the end of September and the beginning of October. In October, the analysis of this data has started and will continue until the presentation of the first Proof of Concept (PoC) in November.

The development of the PoC lays the groundwork for the next development steps to improve and further implement the components needed to meet the requirements described in the User Stories. From November on, the requirements for each User Story as described in D2.2 will be held against the dataset available through the testbed. Short term next steps will include:

- Execution of area risk profiling and conclusions for insurance product pricing (February 2021).
- Seasonal climate forecasts (March 2021).
- Monitoring of satellite derived indices (throughout vegetation period).
- Testing of weather/risk alerts (throughout vegetation period).

### 2.14.7 Conclusions - Issues and Barriers

Based on the work done so far, we are aware of the following challenges:

- Provision of UNINOVA IT infrastructure to run the testbed as foreseen in WP6 to enable enough computing capabilities to run the Weather Intelligence Engine
- Receiving appropriate and high-quality dataset from insurance company for the PoC and ongoing activities
- Identifying the right correlations between the data provided via the testbed and the dataset for the respective AOI in order to draw the right conclusions for the area risk profile and hence the insurance pricing for a drought index insurance product

To conclude the status of the pilot site preparation it can be stated that the partners involved in Pilot #14 (AGRO,GEN) are in close contact with Nova as the Testbed provider and are awaiting their notification of a successful set-up of the shared testbed infrastructure in the coming weeks.

Furthermore, a good relationship was established with two agricultural insurance companies which would be able to provide the for this pilot required insurance company data (see Figure 51) for the defined AOI.

Both insurance companies approached are composite insurance companies, hence not only focusing on agricultural insurance. The Agricultural Line of Business (LOB) of insurance companies in most markets is not the most profitable one. On the one hand, the service module developed in Pilot #14 will contribute to exploiting untapped market potential and new/innovative business and product opportunities, on the other hand though, it is difficult to convince the Management and the Underwriting Departments of all benefits.

Therefore, GEN is using its business relationships to directly talk to potential decision makers. To convince those decision makers, GEN has pitched the overall goal of the INFINITECH project together with the objectives of Pilot #14, the structure of the pilot in general terms, data requirements and lastly the benefits in the short, medium, and long term for the pilot user (as defined in the user stories for agricultural insurance companies).

In a next step, decision-makers will be given time for feedback and questions. Afterwards, a meeting together with the Tech-Proxy of Pilot #14 (AGRO) will be organized to dive deeper into the set-up of the service module, the capabilities of the module to provide additional data and to discuss the data requirements to be derived from the insurance company.

This process is essential for reaching out to potential pilot users in order to test and evaluate the added value of the service module (based on defined user requirements) developed for the specific business processes in agricultural insurance.

## 2.15 Pilot #15 Inter-Banking Open Pilot

The Inter-Banking Open pilot, as explicated by its name, is the result of an Open Call to shared business pains among several Banks, and its objective is to develop a solution that could address and tackle such pains in a pre-competitive environment. Due to its composition, the pilot is strongly market-driven and aims to implement the prototype of a solution based on Machine Learning and Natural Language Understanding paradigms.

This prototype will start from the analysis of a subset of process operating documents to attempt the classification of the information contained in them with respect to the ABI Lab taxonomy, used by Italian banks to build their business glossary and in general to support the Enterprise Architecture Modelling.

ABI Lab is the Banking Research and Innovation Centre founded and promoted by the Italian Banking Association (ABI). Through research and advocacy, ABI Lab promotes innovation as a mean of growth and reinforcement of the banking system. To support digital transformation, ABI Lab has created the AI Hub, a centre of excellence to discuss over the AI application in the banking and financial sector.

Within the AI Hub, the objective of the pilot is to promote the development of a common use case, which will involve different banks through a shared research approach. The use case will be developed following two steps, as described in Figure 52.



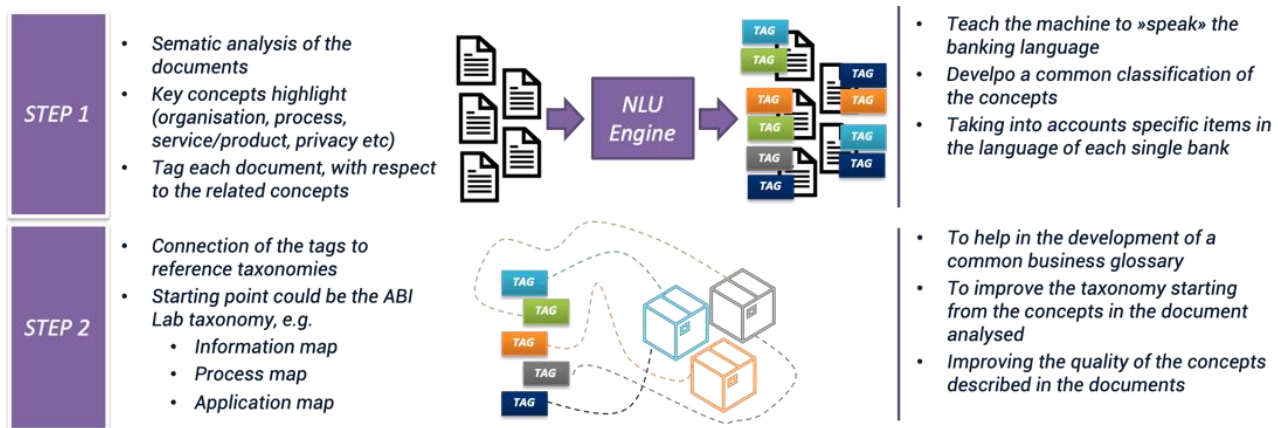


Figure 52: Pilot #15 Steps and main objectives

A high level view of the functional architecture is described below:

- A data storage layer, including tools and infrastructures aimed at data collection from different sources and in different formats, and their storage.
- A data ingestion/preparation layer, including technical components aimed at normalising and aggregating the data needed for this specific analytical purpose, preparing the information to be processed by the Machine Learning tools.
- A machine learning engine layer, including Natural Language Understanding algorithms, opportunely configured for the use case purposes.

This pilot will allow the screening of extensive documentation in real time. This will be a starting point for the optimization of solutions that every single bank can possibly adopt and adapt in their own context. The pilot will involve a community of banks, which will:

- Provide data-set related to internal documentation
- Provide information and addressing issues around the usage of common taxonomy or glossaries to build a classification ad analysis model
- Participate to the requirement identification and service evaluation

The development (and also the training plans for the AI models) will be driven by ABI Lab, supported by the members of the AI Hub community.

The banks will be the final users, keeping into consideration that the objective of the pilot is to develop an experimental prototype that will be the object of further analysis by the participant stakeholders.

### 2.15.1 Technological components and Services

The main objective is to build an AI tool able to read the internal documents of a bank to highlight the main concepts and compare them with reference taxonomies to build a common business glossary.

Technological components and services will be defined according to the pilot objectives.

### 2.15.2 Data sets status

Participating banks will provide data related to internal documentation (for example internal policies, internal circulars, operating guides, user manuals, etc.).

Considering that the Pilot will be based on a machine learning algorithm, the banks participating to the pilot will provide an adequate volume of data (internal documents and glossaries) to support the training and evaluation phase.

### 2.15.3 Testbed

The technical and development aspects, in particular within the dedicated testbed, will be supported by GFT and HPE. The pilot will be hosted and deployed on the Testbed blueprint that will be developed accordingly to the pilot requirements.

### 2.15.4 Others non-technical requirements

Not relevant at this stage.

### 2.15.5 Implementation of a first Proof of Concept

The implementation of the first Proof of Concept will support the user stories defined in D2.2.

### 2.15.6 Next steps and timeline

The pilot started later in respect to the others due to the entering of ABILAB during the project and essence of the pilot itself. Indeed, the decision of the Pilot was the consequence of an Open Call held within the AI Hub, indeed finding an agreement with several Italian Banks participating at the Hub and interested in developing a shared pilot. The business pains of the banks, together with a methodological approach of categorization and classification of the different possible solutions proposed, enabled ABILAB to understand and decide the scope of the pilot, being in pre-competitive environment, proposed as an innovative solution shareable and adaptable on all the Banks.

About the timeline, it is foreseen to conclude the assessment phase, the analysis of the requirements and the study of the alternatives in terms of model, technology, and the collection by the beginning of 2021.

The development phase will start in parallel, aiming to have a POC and a demo available for April. The implementation and refinement phase will follow.

### 2.15.7 Conclusions - Issues and Barriers

This pilot will allow the screening of extensive documentation in real time. This will be a starting point for the optimization of solutions that every single bank can possibly adopt and adapt in their own context. Indeed, the scope of the pilot could arise some foreseen challenges, mentioned below:

- Put together multiple banking, technical and academic stakeholder to achieve shared objectives.
- Harmonization of semantic representational models in the context of financial services.
- Exploitation of data asset.

### 3 Conclusions

During the first year of the project, pilot focused on use cases definition, requirements identification, reference architecture, and corresponding deliverables. Great effort from all pilots covering requests coming from different partners and workpackages; working as a whole. Communications have been crucial to organise and progress in a proper way. Last months the efforts of contributing to deliverables have been running on parallel with development phase, supporting a very specific/quick sprint from July to November (last general assembly). This work led into a Proof of Concept (PoC) for all pilots (with the exceptions explained at the introduction) that summarizes developments and achievements. This PoC also refined the targets of each pilot, whilst helped them to identify new requirements and envision possible constraints and issues. This way, every pilot can work on an improved and more fruitful outcomes within its cluster. There are however few pilots that are lagging behind in terms of their specification and implementation, as a result of late inclusion of some partners in the consortium. The following pilot provides an overview of the status of the various pilots and illustrates that most pilots have managed to implement an initial proof-of-concept and demonstrator:

Pilot Theme	Implementation Status (As of November 2020)
Invoices Processing Platform for a more Sustainable Banking Industry	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Partial Implementation of PoC.</li> </ul>
Real-time risk assessment in Investment Banking	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>
Open Inter-Banking Pilot	<ul style="list-style-type: none"> <li>Pilot Defined and Planned.</li> <li>Stakeholders Engagement Planned.</li> </ul>
Collaborative Customer-centric Data Analytics for Financial Services	<ul style="list-style-type: none"> <li>Pilot still at Specification Stage due to late inclusion of key stakeholders in the project.</li> </ul>
Personalised Portfolio Management (“Why Private Banking cannot be for everyone?”)	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>
Business Financial Management (BFM) tools delivering a Smart Business Advise	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>
Personalized Closed-Loop Investment Portfolio Management for Retail Customers	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>
Avoiding Financial Crimes	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>
Platform for AML supervision	<ul style="list-style-type: none"> <li>Stakeholders Mobilized.</li> <li>Architecture Finalized.</li> <li>Implementation of Initial Integrated PoC.</li> </ul>

Real-time cybersecurity analytics on financial transactions' data	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>
Analysing Blockchain Transaction Graphs for Fraudulent Activities	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>
Personalized insurance products based on IoT connected vehicles	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>
Real World Data for Novel Health-Insurance products	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>
Alternative and automated insurance risk selection and insurance product recommendation for SME's	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>
Big Data and IoT for the Agricultural Insurance Products Industry	<ul style="list-style-type: none"> <li>• Stakeholders Mobilized.</li> <li>• Architecture Finalized.</li> <li>• Implementation of Initial Integrated PoC.</li> </ul>

Table 18 High Level Overview of Pilots' Implementation Status

The successful implementation of Proof-of-Concepts for most of the INFINITECH pilots provides evidence of progress and readiness for the pilots, while at the same time manifesting the collaborative efforts and the synergies between the INFINITECH partners.

So far, pilot's development has been running pretty in parallel, because of time restrictions, with the technological WPs. Not showing, in general, the technological match between pilots needs and INFINITECH provided technologies. This also happened because of INFINITECH technologies have been involved in a first definition process. The work done to contribute to this report and putting together all PoCs helped to break these silos between pilots and share technologies and architecture components. In the coming months, Pilots results will be able to show, use and integrate the results provided by technological INFINITECH workpackages. This will happen with the INFINITECH technologies more defined (deliverables in WP3, WP4 and WP5), and the testbeds and sandboxes ready to support this integration.

WP6 testbeds and sandboxes will unify the way that pilots are prepared, from an infrastructure and deployment perspective, to set the base for similar use cases, or stakeholders with similar needs, to try INFINITECH technologies. Actually, D2.13 Reference Architecture settles the bases for multi-layer architecture, with different components that can be plugged and combined. This approach is been followed by all pilots and it will finish with the all the corresponding testbeds deployment. This way, Kubernetes and Docker's orchestration framework will demonstrate these multi-layer-plugable approach defended by INFINITECH. Pilots already started to work (some fully prepared) with Docker containers for a smooth transition and implementation of sandboxes.

This first phase could be summarized with pilots focusing on data collection and preparation and deployment of first set of components. Data capture, filtering, homogenization (following the Reference Architecture) is already there, and it is starting to work. In the coming phase, these components will be complemented with the results of INFINITECH technologies and services (W3-WP5) and the creation of sandboxes (WP6) that fill finalize a common way of working. While finishing this deliverable, a testing infrastructure is been put in

place. Pilots will have an infrastructure to manage their software components, CI/CD tools for deployment and an environment to create/use blueprints for their architectures. These blueprints will help replicability of similar scenarios and needs, e.g. how to get/inject data through a data pipeline into a LeanXcale database for later analysis.

Cluster 1 comprehends different pilots linked mainly by services and risk assessment purposes. The overall development and deployment of the pilots is proceeding as planned, with an exception for Pilot#15 as highlighted in various parts of the deliverable. Generally, the requirements and development phase for the first two pilots is at an advanced state, on one hand, having already implemented the PoC, deployed on a on-premise cloud-based testbed, on the other hand, having already implemented the PoC that will be deployed on the shared testbed hosted by NOVA. Instead, Pilot#15 will be deployed in the testbed blueprint: indeed Cluster 1 provides a comprehensive view of pilots' deployment by exploiting the three different "typologies" of testbeds established by the INFINITECH project. Overall, the development and training of machine learning applications is proceeding, enhancing the innovative components of the pilots. To conclude, Pilot#1 already planned to perform relevant activities in stakeholders' engagement, demonstrating the maturity and readiness of such pilot, whereas Pilot#2 is a clear example of how INFINITECH technologies can be exploited to develop a trading-based risk assessment use case.

Cluster 2 of Pilots that are related on Personalized Retail and Investment Banking Services, based on the progress until now, they are progressing following their initial plans (except Pilot #3 that is in the process of redefining its scope). The majority are in the process of building the ground for each pilot, which includes mainly the AI power tools that will be used as basis for the final deployment. Most of the pilots either established or in the process of testbed deployment and now based on the relative blueprint definition will start working towards to INFINITECH way of deployment. Even though the main activities already reported are mainly focus on the technical site, the actual goal for each pilot focus on providing technologies that will improve the financial health of individuals and SMEs, either through better and personalized investment propositions or better financial management tools.

At cluster 3 level, we can summarize the current status of implementation of pilots at a general good and promising point. Each of these pilots implemented a PoC initial prototype, provided initial bunch of data and a testbed installation to implement the first serviced and develop the need it technology. First data analytics components, risk calculation engines, complex search services, and user interfaces (for risk assessments), have been developed. Financial Information synthetic data are currently used in research environment at JSI. Blockchain technologies are started demonstration to provide more secured and trusted transactions systems, facing the difficulties of a so high computational demand derived from these technologies: transaction dataset preparation, huge (scalable) transaction graph analysis and visualization tools. A case apart is represented by Pilot #7, because of a change in pilot partners, and the subsequent need for updated specifications, as well as for the redefinition of the pilot according to the new partners.

Cluster 4 is focused on different insurance services customisation, by exploiting real world data collected from users through different AI powered technologies that evaluate the insured client's behaviour and his/her associated potential risks. This first stage on the cluster pilots' development analysed the different available data sources, identifying which are relevant for the use cases to be played and built all the mechanisms needed to gather, curate and homogenise these identified datasets. In parallel, the infrastructure to collect, store and classify the information has been defined and implemented, so, aligned with INFINITECH development and deployment guidelines, the second stage, which will design, build and run all the novel ML models. These ML/DL models will be based on cutting-edge AI technologies and will be specifically created to solve the particularities of each scenario. In turn, they will be the key component on the final new services to be offered to insurance companies and insured clients.

Cluster 5 is focused on customized and configurable insurance products based on non-traditional data sources and not obtained directly from the insured subjects. The objective is to obtain a better determination of the insured risk, the insured enterprises and agricultural sector. On the one hand, to offer a more adjusted and personalised insurance and on the other hand to speed up the payment of the compensation. The technologies that will be used are based on Machine Learning and AI on large amounts of data obtained from

sources both in text format and in satellite images. The process is composed into three phases, the determination of the relevant sources to provide data to the models and their homogenization for processing. The second is the management of the data within the reference architecture established in the lines of INFINITECH. Finally, ML/DL models will be based on cutting-edge AI technologies and will be specifically created to solve the particularities of each scenario.

Finally, most pilots are now focused on technological developments and not so focused on Business Processes and Stakeholders Involvements:

- **Business Process Change and Innovation:** What is the system changing in the business? How things are done today and how they will be done after INFINITECH?
- **Stakeholders' Involvement:** Who is involved from the business side? Who are the end-users and how they are involved in the pilots? Are there stakeholders' workshops planned to evaluate the pilot systems? How many participants are expected, when they will be scheduled? Do we need to train some users to use the system?

These aspects, together with continue developing and implementing the pilots, will be the key that will guide pilots in the following months. Following to this deliverable, 5 more are coming (one per cluster) where more details about the evolution, technological alignment and business processes will be covered.