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

# D2.10 -Initial Specification of Testbeds, Data Assets and APIs - II

| | |
|---|---|
| Title | D2.10 - Initial Specification of Testbeds, Data Assets and APIs - II |
| Revision Number | 3.0 |
| Task reference | T2.5 |
| Lead Beneficiary | CP |
| Responsible | Marinos Xynarianos |
| Partners | AKTIF BOI BPFI CP CPH ENG FBK GFT GLA JRC NBG NUIG PI PRIVE RB UNIC UNP WEA |
| Deliverable Type | Report |
| Dissemination Level | Public (PU) |
| Due Date | 2021-03-31 [M18] |
| Delivered Date | 2021-03-31 |
| Internal Reviewers | LXS ASSEN |
| Quality Assurance | CCA |
| Acceptance | WP Leader Accepted and/or 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 | Role[1] | Author(s)[2] |
|---|---|---|
| CP | Lead Beneficiary | Marinos Xynarianos |
| AKTIF | Contributor | |
| BOI | Contributor | |
| BPFI | Contributor | |
| CPH | Contributor | |
| ENG | Contributor | |
| FBK | Contributor | |
| GFT | Contributor | |
| GLA | Contributor | |
| JRC | Contributor | |
| NBG | Contributor | |
| NUIG | Contributor | |
| PI | Contributor | |
| PRIVE | Contributor | |
| RB | Contributor | |
| UNIC | Contributor | |
| UNP | Contributor | |
| WEA | Contributor | |
| LXS | Reviewer | Patricio Martinez |
| ASSEN | Reviewer | Ilesh Dattani |
| CCA | Quality Assurance | Paul Lefrere |

[1] Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

[2] Can be left void

# Revision History

| Version | Date | Partner(s) | Description |
|---|---|---|---|
| 0.1 | 2021-01-31 | CP | ToC Version 1.0 |
| 0.2 | 2021-02-10 | CP | ToC Version 2.0 |
| 0.3 | 2021-02-10 | CP | Initial Version with new contents |
| 0.4 | 2021-02-25 | CP | Modifications based on 6.7/6.10 detail specification |
| 1.0 | 2021-03-08 | CP | First Version for Internal Review |
| 1.1 | 2021-03-09 | CP, LXS, ASSEN | Second Version for Internal Review including comments from reviewers |
| 1.2 | 2021-03-12 | CP, LXS, ASSEN | Third Version including 2nd Internal Review comments from reviewers |
| 2.0 | 2021-03-15 | CP, CCA | Version for Quality Assurance |
| 3.0 | 2021-03-31 | CP | Version for Submission |

# Executive Summary

The goal of task T2.5 “Open Banking APIs, Testbeds and Data Assets Specifications” and this deliverable is to provide the specification of the Open APIs that will be implemented as part of the INFINITECH-RA, the specification of the project’s data assets and testbeds. An initial version of these specifications has been provided in D2.9 while this second version of the deliverable updates the findings documented there.

Furthermore, this version of the deliverable provides updates on the specification of the testbeds of the project, in terms of functionalities, data assets, as well as in terms of the regulatory tools that they will offer. The document aims to provide updated specifications for the testbeds that will be implemented as part of the INFINITECH Reference Architecture (INFINITECH-RA) already documented in other deliverables, regarding the relevant functionalities, Open APIs, data assets, that will be used and developed from the relevant pilots. It also includes all updated specifications for the testbeds, as well as the regulatory tools that will be used for INFINITECH project Pilots, and updates to the datasets that will be used.

In addition to what has already been provided in the previous version, this document provides more details on the Open APIs that will be implemented and deployed from INFINITECH Pilots and relevant partners, regarding the specifications, high level design principles and technologies that will be used.

This deliverable is the latest version of two deliverables which are meant to provide the outcome of task T2.5, intending to report any updated information available to the initial specifications for:

- The testbeds that will be implemented as part of the project for hosting the relevant pilots;
- The data assets that will be used by each pilot and the respective sandboxes that will be hosted from each testbed, based on the relevant use cases and stakeholders' requirements;
- The Open APIs that will be available for accessing the data assets or will be developed and be available to be used for each of the pilot sandboxes and testbeds, that may be reusable from other INFINITECH Pilots;
- How the tools accessed via the Open APIs will take into account the Regulatory Compliance Tools that will be used in order to meet the requirements of certain regulations (e.g., GDPR, PSD2, MIFID, 4MLD, etc.).

# Table of Contents

# List of Figures

# Abbreviations

| | |
|---|---|
| API | Application Programming Interface |
| DB | Data Base |
| DL | Deep Learning |
| DLT | Distributed ledger technology |
| DoA | Description of Action (also called DoW, description of Work, PART A of Grant Agreement) |
| GDPR | General Data Protection Regulation |
| HTTP | Hypertext Transfer Protocol |
| IoT | Internet of Things |
| KYC | Know Your Customer |
| KYB | Know Your Business |
| MiFID | Markets in Financial Instruments Directive |
| MiFIR | Markets in Financial Instruments and Amending Regulation |
| ML | Machine Learning |
| NDA | Non-Disclosure Agreement |
| NIS | Network and Information Systems |
| OES | Operators of Essential Services |
| PAN | Primary Account Number |
| PaaS | Platform as a Service |
| PCI DSS | Payment Card Industry Data Security Standard |
| PIA | Privacy Impact Assessment |
| PSD2 | Payment Service Directive 2 |
| PSP | Payment Service Provider |
| PSU | Payment Service User |
| P2PP | Peer-to-Peer Payment |
| QTSP | Qualified Trust Service Provider |
| RA | Reference Architecture |
| REST | Representational state transfer |
| RTS | Regulatory Technical Standard |
| SCA | Strong Customer Authentication |
| SHARP | Smart, Holistic, Autonomy, Personalized and Regulatory Compliance |
| SME | Small and Medium-sized Enterprises |

| | |
|---|---|
| SA | Supervisory Authority |
| SECaaS | Security-as-a- Service |
| TI | Threat Intelligence |
| VDIH | Virtualized Digital Innovation Hub |
| XML | Extensible Markup Language |
| 3DS | Three-Domain Secure |
| 4MLD | Fourth Money Laundering Directive |

# 1 Introduction

Task 2.5 of the INFINITECH Project provides the specification of the advanced experimentation infrastructures (testbeds & sandboxes), which shall provide access to resources for application development and experimentation of BigData, IoT and AI-based innovations, as well as the specifications of related data assets, regulatory tools, libraries of ML/DL algorithms, Open APIs, that will be implemented as part of the INFINITECH-RA. Such experimentation infrastructures (testbeds & sandboxes) should be available in various configurations, based on the deployment of the relevant technical building blocks.

## 1.1 Objective of the Deliverable

This deliverable updates the information of D2.9 with respect to changes regarding the specifications of the advanced experimentation infrastructures (testbeds & sandboxes) that will be implemented as part of the INFINITECH-RA, taking into account the functionalities of the technological building blocks that will be developed, whose usage has already been identified their usage within the INFINITECH project, along with the relevant data assets and needs for compliance to specific regulations (e.g., GDPR, PSD2, MiFiD, 4MLD). Also, this deliverable provides an update and extension of the information specification of Open APIs and other resources that will be available or will be developed, based also on the relevant input received during the progress of WP5 ("Data Analytics Enablers for Financial and Insurance Services"), reported in its corresponding first wave of deliverables.

The overall main areas of the updated information included in this deliverable are:

- **Updated guidelines** of the deployment of **advanced experimentation infrastructures (testbeds & sandboxes)** that will be implemented for hosting each pilot related to BigData, IoT and AI-based innovations, based on the required significant testing and validation efforts
- **Updated specifications** of the **data assets** that will be selected, flexibly configured and made available to the relevant **sandboxes** over these testbeds
- **Open APIs, and regulatory tools** needed for testing/ validating the use cases in the target pilots in the finance/insurance sector.

## 1.1. Insights from other Tasks and Deliverables

The deliverable will utilize pilot descriptions and user stories developed in Task 2.1 and available through deliverable D2.2 (PU) under the functional services view from all pilots as described in D2.4 (PU). Moreover, partners involved in Task 2.3 also contributed by providing insights with respect to the INFINITECH core technologies that will be included in deliverable D2.6. Finally, Task 2.4 deliverable D2.8 (PU) defines the regulatory and compliance requirements for each pilot (e.g., GDPR, PSD2, MiFiD, 4MLD, or others) and deliverable D2.13 as part of Task 2.7, will be used as input (see Figure 1 below):

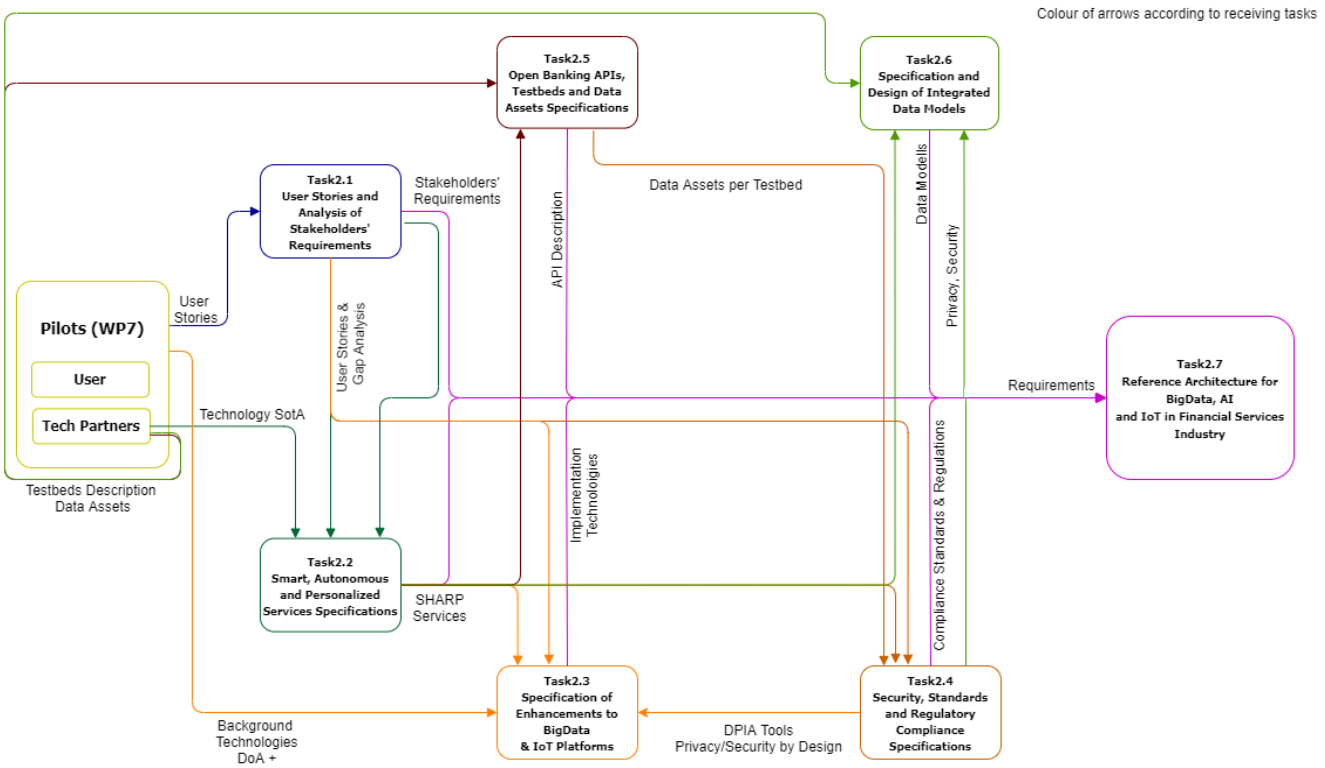


Figure 1 – Schema of the links among Tasks

Also, the deliverable relies on feedback requested from all pilots to contribute regarding data assets that will be used for each pilot and the relevant testbed host.

Datasets for algorithm-training & evaluation that will be used from various pilots are referenced based on Task 5.1 deliverable 5.13 and also Open APIs specifications for accessing the added-value analytics functionalities of INFINITECH that are included in deliverable 5.10 of Task 5.5.

## 1.2. Structure

This deliverable has four main sections:

- Chapter 1 is the introduction to the deliverable and includes the description of the objective, insights from other tasks and deliverables and the structure
- Chapter 2 describes the methodology followed for the collection of the relevant information from all INFINITECH Partners included in the deliverable
- Chapter 3 contains the Initial Specifications regarding:
    - Testbeds & Sandboxes,
    - Data Assets,
    - Open APIs and
    - Regulatory Tools

that will be implemented as part of all INFINITECH Pilots' execution.

- Chapter 4 reports conclusions of the document
- Chapter 5 contains the reference documentation used in this deliverable
- Chapter 6 contains Appendix 6.1 – Data Assets Details

# 2 Methodology

The major source of insights of deliverable D2.10 is the INFINITECH Pilots and their contributions to other deliverables of WP2:

- D2.2 User Stories and Stakeholders' Requirements – II (PU)
- D2.4 Reference Scenarios and Use Cases – II (PU)
- D2.6 Specifications of INFINITECH Technologies – II
- D2.8 Security and Regulatory Compliance Specifications – II (PU)

The Data Assets Specifications are already gathered and reported on in the previous version of the deliverable (D2.9) based on a related spreadsheet that all INFINITECH Pilots contributed to, and in this version includes all the updated relevant information, as described in Paragraph 6.1.

Testbed Initial Specifications are updated based on all INFINITECH Pilots' contributions for deliverables D2.6 and D2.13 - INFINITECH Reference Architecture – I, that is also the basis for all Pilots' implementation.

The results of those contributions are explained in the following sections.

# 3 Testbeds, Data Assets, Open APIs, updated Specifications for the Pilots

## 3.1 Testbed Specification

### 3.1.1 Testbed Definition

A **testbed** [1] is a combination of the hardware and software environment on which tests will be executed. It includes specific hardware, software, operating system, network configuration, test terminals, the software product or solution under test, other system software and application software that may be required for administrative and monitoring functions (see Figure 2)[3].

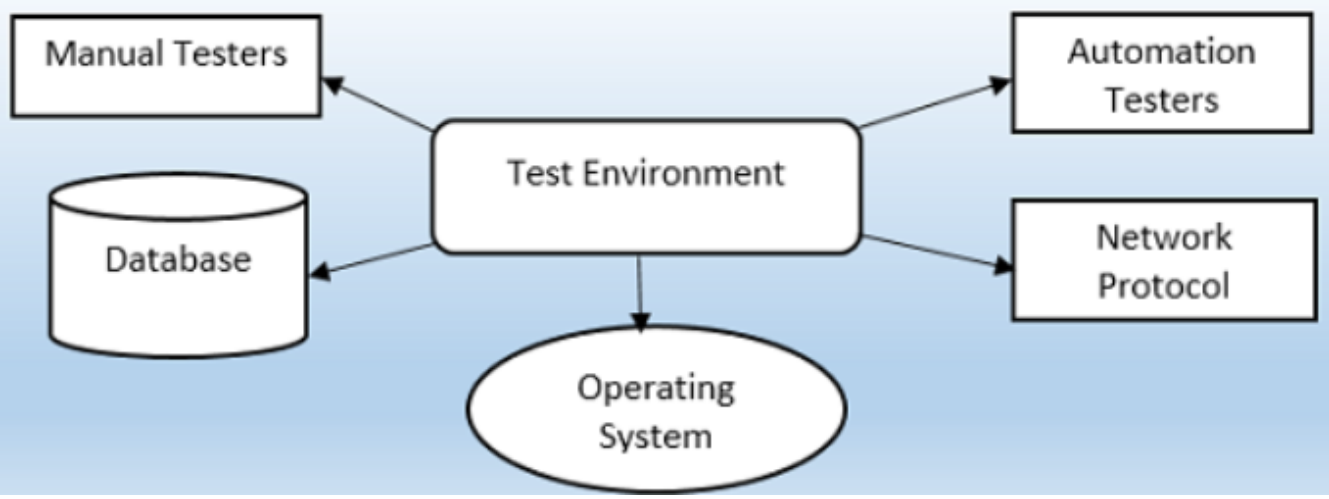


Figure 2 – A composition of various components forming a "Test Bed"

For INFINITECH Project testbeds will host the advanced experimentation infrastructure, which shall provide access to resources for application development and experimentation, such as datasets, regulatory tools, libraries of ML/DL algorithms, Open APIs and more, that will be developed within the context of the project.

#### 3.1.1.1 Design considerations for Test Bed Environment:

Testbed design & implementation for INFINITECH project Pilots, required to address the following Challenges:

- **Setting up a Test Bed on premise or in cloud:** a Testbed's environment could be installed at various locations, it can be installed *either on premise or be cloud based*, which means that it may not be deployed at a centralized location. Given the diversity of the modern environment, where most companies are using cloud-based environments for testing (to reduce the relevant administration and

[3] https://www.professionalqa.com/test-bed

maintenance cost), cloud implementation is one of the more-preferable setups. For some organizations, depending of the relevant legislation requirements applied to security of data and environment, on-premise will be an alternative that may also be considered

- **Share usage among Teams:** In the context of Pilots' implementation software testing, developers, testers and business analysts from the core team will have to perform testing using the same testbed environment that a Pilot's final deployment will make available. Based on this assumption, we can specify the relevant resources regarding software and hardware requirements that need to be taken into consideration. This requirement can be accommodated from testbeds implementing different sandbox environments in each testbed. Also, usage of related software that allows the differentiation of shared resources like containerization engines must be considered as mandatory.
- **Complex Test Configuration:** Complex testing environment for specific pilots that have interconnections with various external resources, also is another critical point that will have to be required to be addressed from each testbed configuration, either from the hardware and software resources, or the required DevOps resources.
- **Time Consuming:** Testbed deployment for each pilot may vary, and may require various test configurations depending on their complexity, so this is another factor that may also affect the availability of the testbed and the relevant sandboxes that each pilot will finally deploy.

#### 3.1.1.2 Environmental Configuration:

For the INFINITECH Project, the Pilots' Testbed instantiation will contain some of the following types of resources (either infrastructure, platform or service and data level of resources):

a) **Cloud specific** infrastructure and **on-premise** (depending on the specific requirements of each Pilot) (e.g., Amazon Web Services (AWS), Microsoft Azure, HETZNER Cloud, etc.)

b) **Application containerization (app containerization):** an OS-level virtualization method used to deploy and run distributed applications without launching an entire virtual machine (VM) for each application (e.g., Kubernetes, Docker, etc.)

c) INFINITECH available **Data Management Technologies**

d) INFINITECH **Big Data AI and ML/DL Tools** developed within the INFINITECH Project

e) **Blockchain Infrastructure** support.

All the above will be provided to all Pilots as part of the INFINITECH Platform in order to achieve the highest level of similarity among Testbeds' relevant sandboxes. More updated details for **INFINITECH Testbeds, Data assets** and relevant **technologies** are available in the next chapter.

More details regarding the Testbeds Implementation are available into the corresponding WP6 deliverables (in particular in D6.1 - Testbeds Status and Upgrades and D6.4 - Tools and Techniques for Tailored Sandboxes and Management of Datasets [2]).

## 3.1.2 Testbed concept of INFINITECH Project

The INFINITECH project is focused on the **development** of **BigData, IoT and AI-based innovations** that usually requires significant testing and validation efforts. This requires testing for regulatory compliance and the optimization of Machine Learning (ML) and Deep Learning (DL) models. As a result, there is the need for advanced **experimentation infrastructures (testbeds & sandboxes)**, which shall provide access to resources for application development and experimentation, such as datasets, regulatory tools, libraries of ML/DL algorithms and Open APIs. Such experimentation infrastructures should be available in appropriate testbeds, based on the deployment of the technical building blocks in various configurations.

The set of hardware resources of each Data Centre (Storage, Compute and Network) will be considered as a **testbed**, as shown in Figure 3 below.

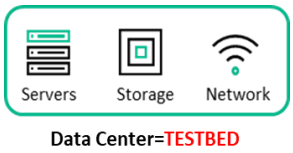


Figure 3 - Testbed

The INFINITECH project will provide **10+2 testbeds** for experimentation, testing and validation of BigData and IoT applications in the financial and insurance sectors, including:

**(i)** **Ten testbeds (10)** that will be established in **incumbent financial organizations** of the consortium

**(ii)** **One testbed (1)** that will be established and made available to Financial/FinTech/InsurTech enterprises of the consortium for their pilots, hosted on the partner **NOVA's Data Center**

**(iii)** **One testbed (1)** that will be provisioned and established in order to support the experimentation of the **INFINITECH blueprint reference** testbed hosted on the AWS (Amazon Web Services) public provider.

Therefore, the 15 pilots will be executed in 10+1+1=12 testbeds (see figure 4 below).

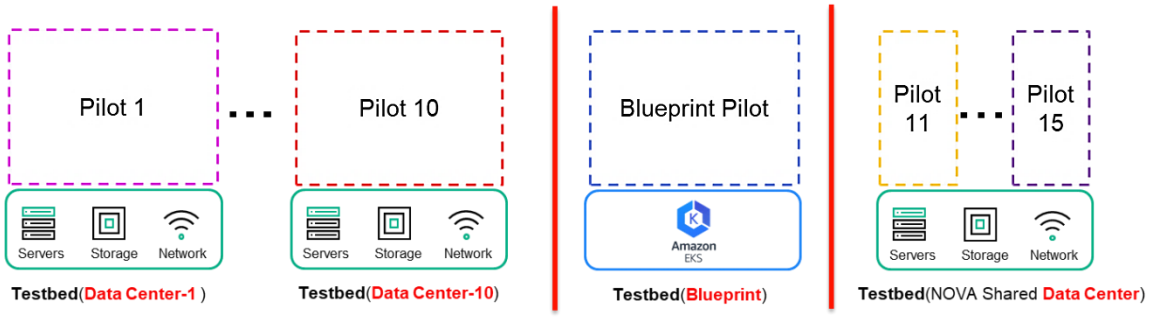


Figure 4 - Pilots vs dedicated shared testbeds

INFINITECH **blueprint reference testbed** deployment aims to cover:

a) the **actual realization** (with full compliance) of the INFINITECH-RA Development and Deployment views, in terms of the concrete specification and realization of the fundamental and target INFINITECH concepts

of Testbeds, Sandboxes and Datasets management, and related tools and techniques for their effective setup and deployment in the INFINITECH pilots and validation scenarios

b) **hosting** the initial and preliminary **Proof of Concept (PoC)** implementation of one or more of the official INFINITECH pilots, depending on the availability of their respective testbed infrastructure

c) fulfil the INFINITECH project goals, regarding setting up the relevant **CI/CD processes** that will be followed, as part of the project development from all partners of the consortium.

Each one of the INFINTECH experimental infrastructures **(testbeds & sandboxes)**, will comprise the following elements:

a) Relevant **Infrastructure** (Storage, Compute and Network) based on **Cloud** services or **on-premise** deployment

b) **Open APIs** (see section 3.2.1 below) for accessing data assets and the INFINITECH data management, analytics, data governance, interoperability and data exchange building blocks

c) **Data assets** for experimentation**,** notably **real anonymized datasets and synthetics/simulated datasets** (see section 3.2 below)

d) **Regulatory tools** for ensuring the compliance of innovative developments with **regulations (e.g., 4MLD, GDPR, PSD2, MiFID2)** (see section 3.4 below)

e) Access to a library of **ML/DL algorithms** for finance and insurance applications that will be developed as part of the INFINITECH project. Specifications of INFINITECH Technologies are available in the deliverable D2.6 - Specifications of INFINITECH Technologies - II [3].

### 3.1.3 Sandboxes concept of INFINITECH Project

A key innovation of INFINITECH is that it will provide the means for provisioning and configuring **tailored sandboxes** over the **project's testbeds**, which will comprise specific data sources, ML/DL algorithms, APIs and regulatory compliance algorithms. **The INFINITECH sandboxes and testbeds** will facilitate innovators in their efforts to produce BigData/ IoT applications that disrupt the sector based on their **SHARP (Smart, Holistic, Autonomy, Personalized and Regulatory Compliance)** properties.

Each INFINITECH pilot will have one or more Use Cases, based on the user stories already described in Deliverable D2.2 - User Stories and Stakeholders' Requirements – II [4], realized by one or more **pilot Apps-Technology Blocks**, each one realized by one or more INFINITECH **microservices**. Each Use Case will be a **Sandbox**.

Each Testbed will host one or more sandbox environments depending on the relevant user stories validation that each pilot decides will require a separate environment for validation of the business requirements.

INFINITECH will provide mechanisms for creating **tailored experimentation environments (i.e., sandboxes)** for **different applications** (e.g., sandboxes for fraud detection, for credit risk assessment, for personalized financial assistance and more), based on **flexible configuration** of the testbed resources.

INFINITECH will enable the development, deployment and business validation of a completely new range of applications that will be characterized by **SHARP (Smart, Holistic, Autonomy, Personalized and Regulatory**

**Compliance)** characteristics. The **INFINITECH pilots** will demonstrate these **SHARP properties** in real-life settings as part of the **project's sandboxes**.

## 3.1.4 INFINITECH Testbeds & Sandboxes concept Diagram

To give further details of the implementation of advanced **experimentation infrastructures (testbeds & sandboxes)** concept within the INFINITECH project, Figure 5 and Figure 6 below describe the main components that each **dedicated** or **shared testbed** will comprise.

**Hardware resources** of each **Data Centre (Storage, Compute and Network)** will be considered as a **Testbed** for each Pilot. Based on the **user cases** that each Pilot will leverage **one or more Sandboxes** will be implemented. Each **Sandbox** will be realized by one or more **Pilot Apps-Technology Blocks**, each one realized by one or more INFINITECH **microservices.** interacting among them through well-known **APIs. For example, a key component for microservices implementation** and deployment is a **Container Orchestrator & Management engine** (e.g., Kubernetes, Docker) that will allow applications or components that will be developed for each sandbox to quickly scale according to the requirements and to easily be updated. The **Open API Gateway** component constitutes the single-entry point for the added-value analytics functionalities and other core offerings of INFINITECH, by enabling the implementation of applications and extensions over them by third-party developers (e.g., participants to hackathons).

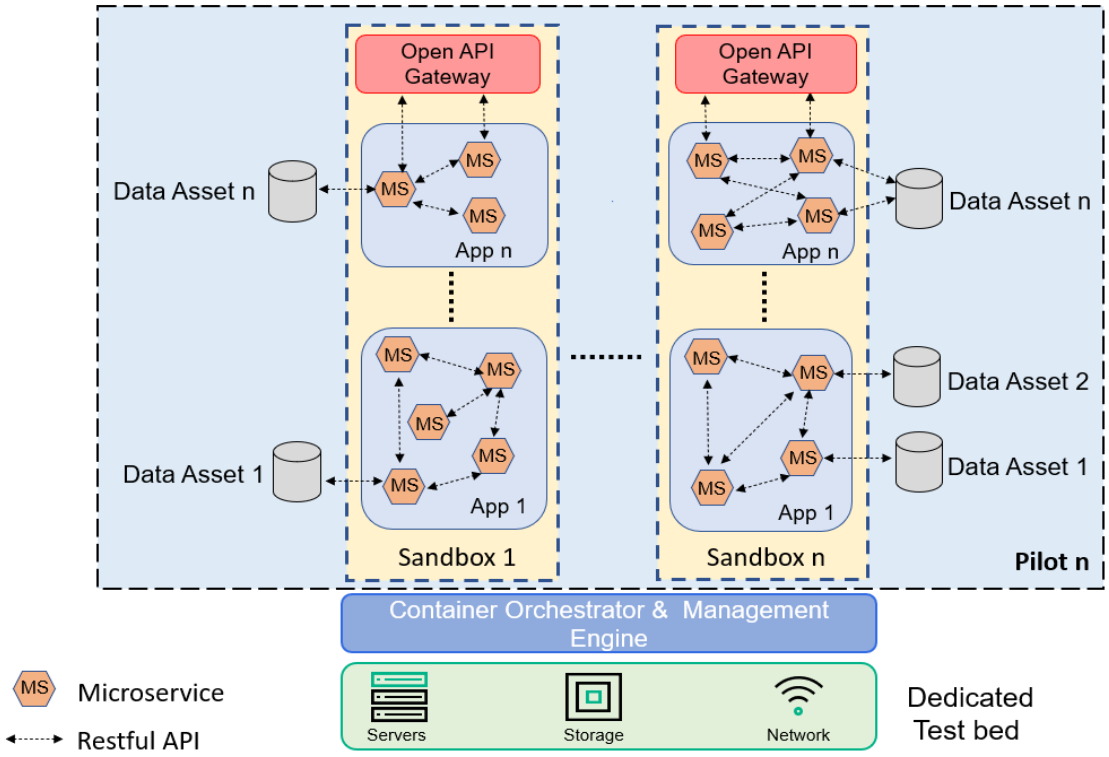


Figure 5 - INFINITECH Dedicated Testbed & Sandboxes Concept Diagram

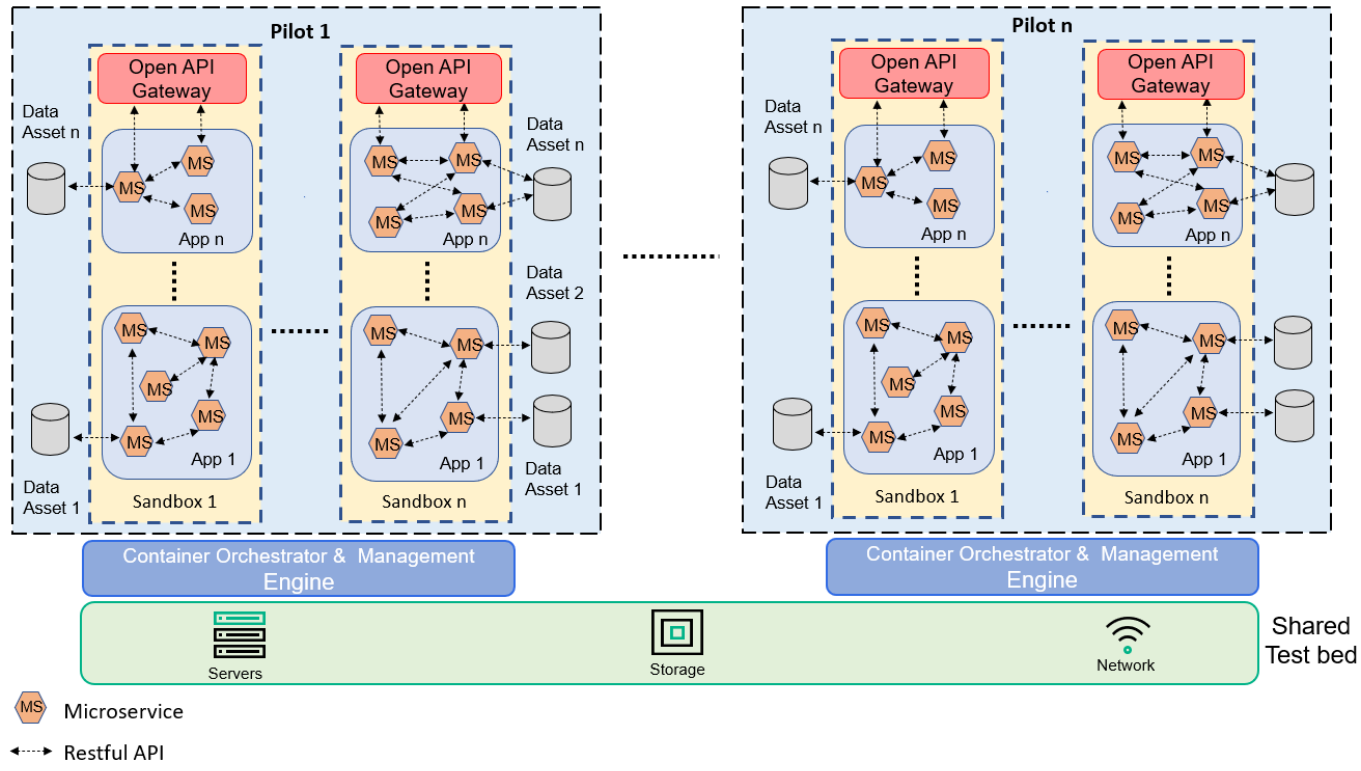


Figure 6 - INFINITECH Shared Testbed & Sandboxes Concept Diagram

More details of the INFINTECH experimental infrastructures (testbeds & sandboxes) are available in deliverable D6.4 - Tools and Techniques for Tailored Sandboxes and Management of Datasets [2]. For example, D6.4 includes more initial technical specifications related to the creation and deployment of the **tailored sandboxes** in the **ten testbeds (10)** that will be established in the **incumbent financial organizations** and will result in a **number of sandboxes** that will be configured according to the continuous integration / DevOps approach, as described in deliverable D6.7 - Sandboxes in Incumbent Testbeds [5].

O**ne testbed** will host the developments of the project's **FinTech/InsurTech firms pilots'** implementations and will result in the deployment of **various sandboxes**, which will provide a first-class opportunity for demonstrating the merits of the **testbed tailoring mechanism**. Initial technical specifications of deployment are available in D6.10 - Sandboxes for FinTech/InsurTech Innovators [6].

Updated technical specifications for all tailored sandboxes and the relevant testbeds mentioned before, will be included in the next versions of the respective deliverables that will be issued at or before the end of the project.

# 3.2 Data Assets Specification

## 3.2.1 Data Assets Definition

Based on the fact that the pilots will validate the INIFINITECH developments from both a technical/technological and a business/economic perspective, they will leverage readily available **BigData datasets** (i.e., data assets) that are available in the organizations of the consortium, **high velocity data from IoT devices** (e.g., connected

cars, medical devices, smart phones), along with **alternative data** from a wide array of **open sources** like **news and social media**.

These data assets include **many millions of customer records**, **billions of customer transactions**, **streams**, **billions of alternative data items** (e.g., news and social media) and more. These data assets will be used by the pilots, while some of them will be also made accessible as part of the project's VDIH Virtual Digital Innovation Hub in order to facilitate the rapid development of BigData/IoT applications in finance/ insurance and to boost the innovation capital of SMEs (including FinTech/InsurTech).

In order to address all Pilot implementations, the INFINITECH-RA has been designed in order to take into account the different types of data sources that each pilot will use. Using specific techniques and tools will allow the tailored sandboxes to enable the data management of diverse environments, mainly focused on data access. Different components that will be deployed inside the sandbox can be integrated in a way that allows them to make use of all the aforementioned data sources.

We can identify six (6) different focuses for data sources of the Data Assets that INFINITECH Pilots will use and have data access:

- **Static data** sources
- **Dynamic data** sources
- **IoT streaming data** sources
- **On-premise data** sources
- **Blockchain data** sources
- **Third party data** sources

With respect to the implementation to support the functionality needed for data management and processing, data governance, details have been given in the corresponding deliverables of WP3, WP4 and WP5.

### 3.2.2 Data Assets Categories

The main categories of the Data assets that will be utilized for INFINITECH Pilots are the following:

a) **Invoice Data**

b) **Financial Market Data (Real Time or History)**

c) **Customer Info Data (Anonymized, Pseudo – Anonymized) – Retail & Corporate (SMEs)**

d) **Customer Behavioral Profile Data**

e) **Customer Risk Analysis Data**

f) **Customer investment Profile Data**

g) **Transactions Detail Data (Retail & Corporate)**

h) **Sentiment analysis Data (News & Social Feeds)**

i) **Financial market price data**

j) **Financial Instruments Characteristics Data**

k) **TARGET2/SEPA transactions Data**

l) **Local or International Business Registries Data**

m) **Bitcoin & Ethereum Blockchain Data**

n) **Black List Data (Individuals, Corporates)**

o) **Traffic & Vehicle IoT Data**

p) **Vehicle & Life Insurance Data**

q) **Health Data**

r) **SMEs Geolocation & Characteristics Data**

s) **Gridded Climate Indices Data (History)**

t) **Earth Observation IoT Data**

These **data assets** will be available through different sources, e.g., Data Lakes, Data Warehouses, Operational Databases, Social & Internet Feeds, Data extraction files in various formats, **IoT Devices Data** (e.g., in the case of the usage-based insurance pilots) and other forms. These data assets will also be the data that will be used to train the algorithms of the project's ML/DL library. Other data may also be used including: **anonymized data** from financial institutes and insurance companies, **open source** and alternative data, as well as relevant **synthetic datasets** that will be produced in order to facilitate training and sharing of the data outside the financial institutions of the consortium.

In Deliverable D5.13 - Datasets for Algorithms Training & Evaluation [7] the datasets are described in more detail, along with the data that will be required for training the algorithms of the project's ML/DL library, that will be evaluated and used in the scope of the project's pilots.

Data assets for each pilot will have to follow the required Security and Compliance specifications already reported in D2.8 - Security and Regulatory Compliance Specifications -II [8] deliverable, using the related tools described in Paragraph 3.4 below.

An updated description of the data assets of each pilot and test-bed along with in-depth details, is described in Paragraph 6.1. More details for the datasets used in each pilot (regarding data sources, data collection, data preparation, etc.) are being included in D7.1 – Report on Pilot Sites Preparation -I [9].

# 3.3 API Specification

## 3.3.1 API Specification Introduction

Nowadays, experimentation and testing of digital finance and FinTech applications takes place within specialized environments that are termed sandboxes. Sandboxes come typically with Open APIs that enable innovators and testers to mimic the characteristics exhibited by the production environment. They can operate on a real-time basis and help simulate responses from all the systems that are involved in delivering the service under test, which can facilitate pilot testing and reduce risks associated with novel technologies such as BigData and AI.

**API (Application Programming Interface)** mainly describes the requirements that govern how **one application** can **communicate** and **interact** with **another**, whether are both in the case that they reside within a single computer via a **mechanism** provided by the **operating system** or communicate **over** an **internal** or **external TCP/IP-based** or **non-TCP/IP-based network.**

An **Open API** [10] iis a **publicly available** application programming interface that allows the owner of a **network-accessible service** to give a universal **access** to **consumers** (organizations, developers or other third parties). **Open** and **public APIs** may be considered as **analogous terms**, but may see a difference in use, in that **Open API**s are **shared** and **accessed freely**, where **public APIs** may have a **more restricted access** concerning the assets and the functionality they share.

INFINTECH will provide, enhance, customize and validate in realistic pilots a range of novel technologies, which will facilitate innovations by financial and insurance organizations based on BigData, IoT and AI. These Pilots will utilize **Open APIs** for experimentation connected to each pilot, but also for supporting innovation beyond the specification pilot requirements (e.g., in the scope of hackathons). The Open APIs will be a customized version of the Open APIs of the INFINITECH technology enablers (i.e., the APIs for Data Management and Analytics).

Also, INFINITECH Project aims to establish a **market platform** and of an EU- **Virtualized Digital Innovation Hub (VDIH)**, which will offer financial/insurance organizations and innovators with a unique blend of innovation management services (e.g., training, consulting, development based on **Open APIs**) that will enable them to innovate with BigData and IoT in the finance sector.

## 3.3.2 API Design Best Practices

**Open APIs** may be **designed** in a variety of different ways, but the main goal of any **Open API architecture** is for the API itself to **be deployed** in a way that can be **easily consumed** and **accessed** by as many different clients as possible. Using **proprietary protocols** or **custom data formats** must be **avoided,** while using **open source technology** and **community-driven standards** is **encouraged,** as one of the **good industry** basic **API design guidelines**.

**Open APIs** might be **universally accessible**, but in **many cases** may be **required** to **restrict access**, **encrypt data** transmission, and leverage **API security** measures, **depending on the** data and **access security implied** by the

application that the APIs are exposed to. For example, usage of **Transport Layer Security (TLS)** can **encrypt data** sent across a network. **Identity authentication** of the caller using **security certificates** prove useful, and any **number** of **back-end authentication mechanisms** can **map users** that have **access** to private or **sensitive data**, or to specific functionality (algorithms, etc.).

The **Open APIs** specification that will be available through the INFINITECH Pilots will be based on the **Open API Initiative (OAI)** [11].

The **OpenAPI Initiative (OAI)** was created by a consortium of forward-looking industry experts who recognize the immense value of **standardizing** on **how APIs** are **described.** As an open governance structure under the Linux Foundation, the OAI is focused on creating, evolving and promoting a vendor neutral description format. The **OpenAPI Specification (OAS)** was originally based on the Swagger Specification, donated by SmartBear Software.

The **OpenAPI Specification (OAS)** defines a standard, programming language-agnostic interface description for REST APIs, which allows both humans and computers to discover and understand the capabilities of a service without requiring access to source code, additional documentation, or inspection of network traffic. When properly defined via OpenAPI, a consumer can understand and interact with the remote service with a minimal amount of implementation logic. Similar to what interface descriptions have done for lower-level programming, the OpenAPI Specification removes guesswork in calling a service.

OpenAPI allows you to define how your REST API works, in a way that can be easily consumed by both humans and machines. It serves as a contract that specifies how a consumer can use the API and what responses you can expect.

**OpenAPI v3.0** was released in July 2017, by the **OpenAPI Initiative**, a consortium of member companies who want to standardize how REST APIs are described. There are various other approaches to API description:

- **OpenAPI v2**, (formerly known as **Swagger v2.0** [12]) is still widely used, but increasingly being replaced by OpenAPI v3.0
- **JSON Schema** [13], very similar to OpenAPI, but able to describe any JSON-like data, not just APIs
- **RAML** [14], the RESTful API Modeling Language, focuses on the planning stage of API design
- **Web Services Description Language (WSDL)** [15]**,** an XML-based interface description language that is used for describing the functionality offered by a web service
- **Web Application Description Language (WADL)** [16], a machine-readable XML description of HTTP-based web services
- **Open Data Protocol (OData)** [17], an open protocol which allows the creation and consumption of queryable and interoperable REST APIs in a simple and standard way. Microsoft initiated OData in 2007. Version 4.0 was standardized at OASIS and in April 2015 OASIS submitted OData v4 and OData JSON Format v4 to ISO/IEC JTC 1 for approval as an international standard

- **RESTful Service Description Language (RSDL)** [18]**,** a machine-and human-readable XML description of HTTP-based web applications (typically REST web services).

While OpenAPI v3.0 is the way forward, each of these alternative formats has tooling associated. Pilots' implementers may find themselves converting between them, especially OpenAPI v2.0, until the tools catch up.

Your API design requires a way to define how the API will be used. The future-thinking approach is to select OpenAPI v3.0 to describe your API [19].

## 3.3.3 Open APIs Specifications for INFINITECH Pilots

One of the INFINTECH experimental infrastructures (testbeds & sandboxes) elements are Open APIs for accessing data assets and the INFINITECH data management, analytics, data governance, interoperability and data exchange building blocks. Testbeds and sandboxes Open APIs concepts are adopted within the INFINITECH project, considering that the INFINITECH-RA technology components are designed leveraging a paradigm based on a **microservices** [19] architecture implementation, with services interacting among them through **REST APIs**, as such will be utilized for the majority of the Pilots' execution.

### 3.3.3.1 Microservices Architecture

In short, the **microservice architectural style** [20] is an approach for developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. These services are built around business capabilities and independently deployable and this is the inspiration for the INFINITECH-RA. The introduction of the **microservices architecture** approach solves the **drawbacks** of the **traditional monolithic architecture** approach by structuring the underlying application and in our case each INFINITECH Pilots, as a collection of microservices that are loosely-coupled, highly maintainable and testable, independently-deployable, organized by business capabilities and can be owned by different development teams or consortium partners. Figure 7[4] below, describe the differences between monolithic and microservices architecture for application development:

[4] https://martinfowler.com/articles/microservices/images/sketch.png

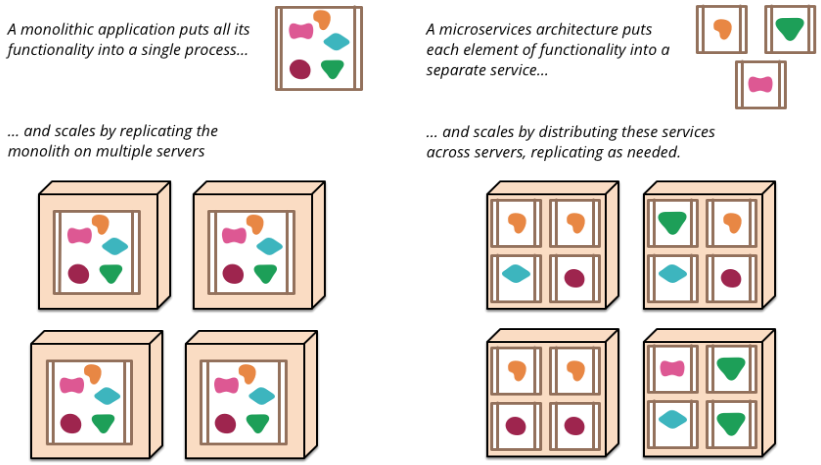


Figure 7 - Monolithic vs. Microservices Application Architecture

### 3.3.3.2 REST APIs

Communication in a microservice architecture [20] is based on lightweight mechanisms, often an HTTP resource API. INFINITECH developments **communication mechanisms** will be based on **REST APIs**. In a nutshell, **REST APIs** (which are a type of web API) involve **requests** and **responses**, similar to the messages that are being exchanged when a user visits a web page. The user make a request to a resource stored on a server, and the server responds with the requested information. The **protocol** used to transport the data is **HTTP**. An example of microservices communication using HTTP and REST is depicted in Figure 8 below:

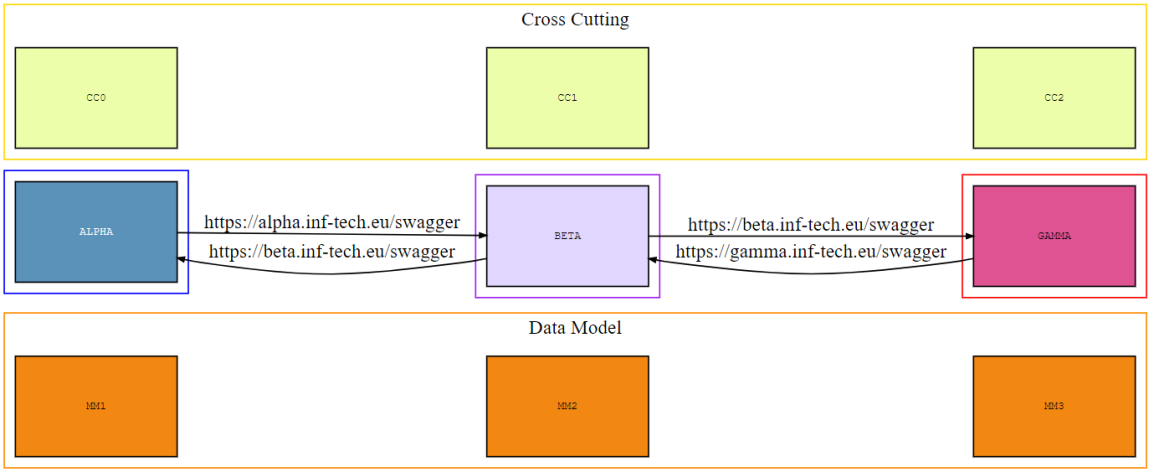


Figure 8 - Microservices Communication example using REST APIs

Find below a list of some of the commonly used Design fundamental of RESTful API:

1. Simple API endpoints
2. Use Nouns instead of Verbs for URIs
3. User Parameters along with URL instead of operational URIs
4. Use sub-resource for data relationships
5. Use the appropriate HTTP method

6. Use proper HTTP error handling
7. Versioning the APIs
8. Other useful operations (Filtering, pagination, Sorting)
9. Decide support request/response formats
10. Use proper/identifiable error messages
11. User standard specification Open API Specification: https://swagger.io/resources/open-api/ world standard for the RESTful API design with the support of other authentication and authorization mechanisms

More details are available in http://bit.ly/3a7ErbY

#### 3.3.3.3 Containerization

**Containerization** [21] has become a major trend in **software development** as the process of packaging together, an application along with its required libraries, frameworks, and configuration files, in order to be able to run in various computing environments efficiently, uniformly and consistently. The technology is provided as an alternative or companion to virtualization and is quickly maturing, resulting in measurable benefits for developers and operations teams as well as overall software infrastructure.

The concepts behind microservices and containerization are similar as both are software development practices that essentially transform applications into collections of smaller services or components, which are portable, scalable, efficient and easier to manage. These characteristics are proposed to be adopted in the INFINTECH experimental infrastructures (testbeds & sandboxes).

Key pillars for the implementation and deployment of a **microservices** based architecture are usage of **containers** technology [22] and its leading open-source container-orchestration solutions, **Kubernetes** [23], and **Docker** [24].

More details about the technical implementation of the INFINTECH experimental infrastructures (testbeds & sandboxes) are available in detail in deliverable D6.4 - Tools and Techniques for Tailored Sandboxes and Management of Datasets [2].

#### 3.3.3.4 API Specifications adoption example

Open APIs will facilitate innovators to access both data assets and algorithms or any other components that each Pilot will develop within INFINITECH through an **Open API Gateway** component, which constitutes the single-entry point for the added-value analytics functionalities and other core offerings of INFINITECH. The INFINITECH **Open API Gateway** is a core ingredient of the INFINITECH Reference Architecture, and one example of implementation that is based on the APIs specifications described above.

**Open API Gateway** design and development specifications details, are described in the deliverable **D5.10 – Data Management Workbench and Open APIs-I** [25]**,** and will be finally provided through the **Market platform and VDIH** that INFINTECH project will develop (WP8) as a repository of all the projects' results, which will be made available to the EU digital finance/insurance communities.

## 3.4 Regulatory Compliance Tools Specifications

Based on D2.8 - Security and Regulatory Compliance Specifications -II [8], all advanced **experimentation infrastructures (testbeds & sandboxes)** for BigData and IoT developments, shall follow the best practices of the relevant data **security frameworks** (i.e., ISO27001, ISO27701, etc.), as well as applicable **regulations** in the financial/insurance sectors (i.e., PSD2, GDPR, 4MLD, MiFIDII/MiFIDR).

The following regulations have been identified as the ones that apply to all the pilots and will be followed based on the relevant datasets' usage and processing:

- **GDPR** for INFINITECH pilot systems that deal with personal data
- **MIFID II** for financial consultancy services
- **PSD2** for online payment platforms
- **AMLD4** for fighting against money laundering and funding for terrorism.

The main types of technologies and the relevant tools that will help to support compliance with the above regulations include:

- For **GDPR Compliance**:
  - Anonymization.
  - Pseudonymization.
  - Privacy dashboards
  - Strong authentication and authorization mechanisms
  - Encryption of data
  - Data Protector Orchestrator.

- For **MIFID II Compliance**:
  - Auditing logs
  - phone call recording
  - email logs
  - strong authentication, preferably multi-factor, and authorization mechanisms

- For **PSD2 Compliance:**
  - Strong multi-factor authentication
  - SIEM (Security Information Event Management) systems

The general philosophy of the INFINITECH project towards regulatory compliance tools is to make use of those tools that are already applied to perform the function of providing support for regulatory compliance. Where existing tools in place are not available or sufficient, users will be provided with new tools.

In general, all testbeds shall follow the best practices regarding **ISO27001**, which are today's generally accepted State of the Art in **Security** measures. Considering **privacy** also the measures related to **ISO 27701** apply for pilots using personal data. Finally, the additional **technical measures GMS-0019** (**Supplier Management**) and **GMS-**

**0020** (**Risk Management**) defined by the **NIST** should be considered at a project level in what concerns the provisioning of infrastructure.

More details for the Regulatory tools (e.g., the relevant Security and Compliance regulations that apply to each pilot, and that will be applied in the relevant Testbed & Sandbox environment) are described in D3.15-Regulatory Compliance Tools – I [26], as well as in the new versions of the deliverable of Task 3.6 of INFINITECH Project.

# 4 Conclusions

This deliverable is the final version of the deliverable which is meant to be provided as the outcome of task T2.5.

The scope of task T2.5 is to describe the initial specification of the advanced experimentation infrastructures (Testbeds & Sandboxes) which shall provide access to resources for application development and experimentation of BigData, IoT and AI-based innovations, as well as the specifications of relevant data assets, regulatory tools, libraries of ML/DL algorithms and Open APIs, that will be implemented as part of the INFINITECH-RA.

In this deliverable, we have included the **final definition** of what advanced experimentation infrastructures (**Testbeds & Sandboxes**) meant for INFINITECH Project and will be implemented, as well as also to describe the main specifications of the relevant **data assets** used for all INFINITECH Pilots that will be hosted.

Within this deliverable, based on the iterative feedback from all Pilots, the final specifications for advanced **experimentation infrastructures (testbeds & sandboxes)**, which shall provide access to resources for application development and experimentation, such as **datasets, regulatory tools**, libraries of ML/DL algorithms, **Open APIs** and more, are described.

Also, the final specifications about the technologies (e.g., BigData/IoT, AI/ML toolkits, HPC infrastructures) will be used to realize and guide the implementation and integration of INFINITECH Pilots, following the **microservices** development, along with the relevant **containerization** technologies that is widely accepted in modern application development.

**Open API** specifications for design and implementation were also included in this version of the deliverable, describing the best of breed specifications that are widely used, based on Open API Initiative and **RESTful** design.

Overall, the INFINITECH Testbeds & Sandboxes, Data assets and Open APIs initial requirements reflect the State of the Art of the application of BigData, IoT and AI in the Financial Services and contribute to the latest trend, that INFINITECH-RA will envisage as part of all pilots and technologies integration.

# 5 References

[1] "Testbed," . [Online]. Available: https://professionalqa.com/test-bed. [Accessed 24 03 2021].

[2] "INFINITECH - D6.4 – Tools and Techniques for Tailored Sandboxes and Management of Datasets," [Online]. Available: https://drive.google.com/file/d/1XMRxmGlPxMvR80iCkIEGEakuvN6xUU_C/view. [Accessed 24 03 2021].

[3] "INFINITECH D2.6 - Specifications of INFINITECH Technologies - II," [Online]. Available: https://app.infinitech-h2020.eu/deliverable/9. [Accessed 24 03 2021].

[4] "INFINITECH - Deliverable D2.2 - User Stories and Stakeholders' Requirements – II," [Online]. Available: https://drive.google.com/file/d/1XGqmRt9WYOUq1CcbbFybjRutSLyVD9qY/view. [Accessed 24 03 2021].

[5] "INFINITECH - D6.7 - Sandboxes in Incumbent Testbeds - I," [Online]. Available: https://drive.google.com/file/d/1bI9rpIYqFAOyiLkt-yB8MTaypcBWAjFC/view. [Accessed 24 03 2021].

[6] "INFINITECH-D6.10-Sandboxes for FinTech and InsuranceTech Innovators-I," [Online]. Available: https://drive.google.com/file/d/1ajM2s8R4cwt2R9znHbwRgmr1inQMO5GW/view. [Accessed 24 03 2021].

[7] "D5.13 - Datasets for Algorithms Training & Evaluation - I," [Online]. Available: https://app.infinitech-h2020.eu/deliverable/135. [Accessed 24 03 2021].

[8] "INFINITECH-D2.8 – Security and Regulatory Compliance Specifications – Version II," [Online]. Available: https://drive.google.com/file/d/1XVZ3RscexhzTYGh1g60fC4MP4e0m_UGY/view. [Accessed 24 03 2021].

[9] "INFINITECH D7.1 – Report on Pilot Sites Preparation - I," [Online]. Available: https://drive.google.com/file/d/1X6d9aGWwl76b7X8w5JB1KtPqaaba8Xml/view. [Accessed 24 03 2021].

[10] T. Cameron McKenzie, "open API (public API)," [Online]. Available: https://searchapparchitecture.techtarget.com/definition/open-API-public-API. [Accessed 24 03 2021].

[11] "OpenAPI Specification," [Online]. Available: http://spec.openapis.org/oas/v3.0.3. [Accessed 24 03 2021].

[12] "OpenAPI Specification v2.0," [Online]. Available: https://swagger.io/specification/v2/. [Accessed 24 03 2021].

[13] "Understanding JSON Schema," [Online]. Available: http://json-schema.org/understanding-json-schema/. [Accessed 24 03 2021].

[14] "RAML Version 1.0: RESTful API Modeling Language," [Online]. Available: https://github.com/raml-org/raml-spec/blob/master/versions/raml-10/raml-10.md/. [Accessed 24 03 2021].

[15] "Web Services Description Language (WSDL) 1.1," [Online]. Available: https://www.w3.org/TR/2001/NOTE-wsdl-20010315. [Accessed 24 03 2021].

[16] "Web Application Description Language," [Online]. Available: https://www.w3.org/Submission/wadl/. [Accessed 24 03 2021].

[17] "OData (Open Data Protocol)," [Online]. Available: https://www.odata.org/. [Accessed 24 03 2021].

[18] "Balisage Paper: RESTful Service Description Language (RSDL)," [Online]. Available: http://www.balisage.net/Proceedings/vol10/html/Robie01/BalisageVol10-Robie01.html. [Accessed 24 03 2021].

[19] "OpenAPI Specification," [Online]. Available: http://spec.openapis.org/oas/v3.0.3. [Accessed 24 03 2021].

[20] "Microservices a definition of this new architectural term," [Online]. Available: https://martinfowler.com/articles/microservices.html#footnote-etymology. [Accessed 24 03 2021].

[21] IBM, "Containerization," [Online]. Available: https://www.ibm.com/cloud/learn/containerization. [Accessed 24 3 2021].

[22] IBM, "Containers," Google, 2018. [Online]. Available: https://www.ibm.com/cloud/learn/containerization. [Accessed 24 03 2021].

[23] Kubernetes, "What is kubernetes?," Kubernetes, . [Online]. Available: https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/. [Accessed 24 03 2021].

[24] "https://docs.docker.com/get-started/overview/," Docker. [Online]. [Accessed 24 03 2021].

[25] "INFINITECH D5.10 – Data Management Workbench and Open APIs – I," [Online]. Available: https://drive.google.com/file/d/1XKaZbSso5A9MwPfwu4lnozgMfUyvCcsk/view. [Accessed 24 03 2021].

[26] "INFINITECH-D3.15-Regulatory Compliance Tools – I-v3.0," [Online]. Available: https://drive.google.com/file/d/1Gm6NhNJZcX2CflQYxO6rAqSGhioa9Ztl/view. [Accessed 24 03 2021].

# 6 Appendices

## 6.1 Data Assets Details

| INFINITECH 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<br>• Dataset TableBank: Table Benchmark for Image based Table Detection and Recognition, 500.000 documents | BANKIA | Confidential data (Notary invoices) & Public Data (TableBank) | No | TBC | PDF/Image/Text | PDF / PNG / TXT | Apache Hadoop HDFS, Elastic Search |
| **Pilot#2 JRC** | Real time financial market data | Price data for the most liquid Forex, Stocks, Stock Indices and Derivatives | JRC | proprietary data from provider | No | Yes | numeric | CSV | MySQL |
| **Pilot#2 JRC** | Derived analysis data | Risk measures, correlation matrices | JRC | open | No | Yes | numeric | CSV | |
| **Pilot#2 JRC** | Existing historical data | Price tick data for the most liquid Forex, Stocks, Stock Indices and Derivatives | JRC | proprietary data from provider | No | Yes | numeric | CSV | MySQL |
| **Pilot#2 JRC** | News articles data | A database of 1.5 billion news articles from 95,654 global news sources will be also used in terms of alternative sources. | JRC | open | No | No | Text | TXT | |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#3, BOI** | Synthetic Customer - Bank A (<100 Records) | | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Applicant - Bank A (<100 Records) | | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Customer - Bank B (<100 Records) | Small dataset of manually produced data to mimic profiles and characteristics of banking customers. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Applicant - Bank B (<100 Records) | Small dataset of manually produced data to mimic profiles and characteristics of banking applicants/prospects. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Customer - Non-Bank Organisation (<100 Records) | Small dataset of manually produced data to mimic profiles and characteristics of a non-banking customer or prospect. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Applicant - Non Bank Organisation (<100 Records) | Small dataset of manually produced data to mimic profiles and characteristics of a non-banking customer or prospect. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Account (<200 Records) | Small dataset of manually produced data to mimic accounts of banking customers. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#3, BOI** | Synthetic Transaction Data (<500 Records) | Small dataset of manually produced data to mimic transactions on accounts of banking customers. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Consent Records (<100 Records) | Small dataset of manually AND/OR POC produced data to mimic history data sharing consent transactions between sharing parties. | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Consent History (<100 Records) | Small dataset of manually AND/OR POC produced data to mimic history of data sharing consent transactions between sharing parties. | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Data Sharing Log (Parties, Metadata and Data Types) (<500 Records) | Small dataset of manually AND/OR POC produced data to mimic data sharing transactions between sharing parties. Maybe stored on Blockchain or as GraphDB (TBD). | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic Customer Input Data. (<1000 Records) | Small dataset of manually AND/OR POC produced data to mimic minimal/basic data sharing application customer/user data (Data TBC). | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#3, BOI** | Optional Storage of Shared Data Set | TBD - Dependent on design. Duplication of records from other datasets. | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Invite to Share Data Set | TBD - Dependent on design. Proposition dataset based on consent to analyze data sharing log. | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Ref. Data. (<1000 Records) (for standardization of data) | Very small dataset of manually produced reference data to support data sharing application(s). | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Customer Linking Data (e.g. iD, Mobile No.) (<1000 Records) | Customer Keys required for data sharing application. | BOI & Pilot 3 | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#3, BOI** | Synthetic or Open Organizational Data e.g. Banks Credentials (TBC) | Very small dataset of manually produced data on organization credentials to support data sharing application(s) including api authentication. | BOI | N/A | N/A | Yes | Text | CSV or JSON | TBD |
| **Pilot#4 PRIVE** | Customer Transaction dataset | Customer securities and cash transactions through their deposit accounts | PRIVE | Confidential data | Yes | TBC | Text/numeric | CSV | TBD |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#4 PRIVE** | Financial market price data | Price data for Stocks, Bonds, Mutual Funds and or other assets like certificates/warrants | PRIVE | Open, partially license agreements with data providers needed | No | No | Text/numeric | TXT | TBD |
| **Pilot#4 PRIVE** | Financial market asset master data | Asset related characteristics (e.g. expiration date, minimum investment amount, asset class breakdowns) | PRIVE | Open, partially license agreements with data providers needed | No | No | Text/numeric | TXT | TBD |
| **Pilot#4 PRIVE** | Customer Risk Profile Data | Customer Risk Profile Data through their account data and profiling, based on B2B customers parameters | PRIVE | Confidential data | Yes | TBC | Text/numeric | CSV | TBD |
| **Pilot#4 PRIVE** | Mutual Fund, ETF and Structured Products Breakdown | Asset Breakdowns based on bank data or market data providers breakdown | PRIVE | Open/Confidential data, partially license agreements with data providers needed | No | TBC | Text/numeric | CSV | TBD |
| **Pilot#4 PRIVE** | Customer Economic Outlook | Customer Economic Outlook data based on questionnaire engine | PRIVE | Confidential data | Yes | TBC | Text/numeric | CSV | TBD |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#4 PRIVE** | Account & Investors data | 19484 accounts for about 15400 investors (live data) 94.407 different securities available; Investors serviced by 309 different advisor companies; Accounts in 28 different custodian banks (Data from 2019) | PRIVE | Confidential data | Yes | No | Text | TXT | TBD |
| **Pilot#4 RB** | News articles data | 5000 articles per day (future) plus 4 million articles (existing) GR News | RB | Open | No | Yes-limited | Text | TXT | Apache Hadoop HDFS, Elastic Search |
| **Pilot#4 RB** | News articles data | Up to 1.2 Million Articles per day from Global News Sources (ENG Language) | RB | Open | No | Yes-limited | Text | TXT | Apache Hadoop HDFS, Elastic Search |
| **Pilot#5b BOC** | Pilot 5b Data Assets Information is Confidential | | | | | | | | |
| **Pilot#6 NBG** | CRM Data | Customer related data like demographics, product ownership and responses to MIFID questionnaires. | NBG | Confidential data | Yes | TBC | Text/numeric | TXT/CSV | Not yet decided |
| **Pilot#6 NBG** | Deposit Account Transactions | Customers' transactions through their deposit accounts | NBG | Confidential data | Yes | TBC | Text/numeric | TXT/CSV | Not yet decided |
| **Pilot#6 NBG** | Cards Transactions | Customers' transactions through their cards | NBG | Confidential data | Yes | TBC | Text/numeric | TXT/CSV | Not yet decided |
| **Pilot#6 NBG** | Instruments Historical Prices | Historical prices of investment instruments | NBG | Open | No | TBC | Text/numeric | TXT/CSV | Not yet decided |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#6 NBG** | Investment Related Transactions | Customers' transactions related to investment products | NBG | Confidential data | Yes | TBC | Text/numeric | TXT/CSV | Not yet decided |
| **Pilot#6 NBG** | Instruments Characteristics | Detailed characteristics for all instruments that will be considered in the pilot, including asset class, currency, ISIN, maturity etc. | NBG | Open | No | TBC | Text/numeric | TXT/CSV | Not yet decided |
| **Pilot#6 NBG** | News feeds & Blogs | Unstructured data collected from public news feeds & Blogs | NBG/RB | Open | No | TBC | Text | TXT/CSV | Not yet decided |
| **Pilot#7 FTS** | Pilot 7 Data Assets Information is Confidential | | | | | | | | |
| **Pilot#8, BOS** | Pilot 8 Data Assets Information is Confidential | | | | | | | | |
| **Pilot#9, BOUN** | Bitcoin Blockchain Data | Bitcoin transfers (send); | Open data | Public blockchain data | Yes | TBC | Text/numeric | TXT | blockchain files |
| **Pilot#9, BOUN** | Ethereum Blockchain Data | Ether transfers (send); ERC20 Token Smart contract transactions (35 popular tokens including stable coins like EURS, GUSD, USDT, TRYB, PAX,TUSD, QCAD, XAUT) | Open data | Public blockchain data | Yes | TBC | Text/numeric | TXT | blockchain files |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#9, AKTIF** | Blacklisted Addresses Data | Small list of blockchain addresses that will be obtained from the Internet by manual search for hacked/fraudulent accounts as well as compiled internally from problematic customers. Bitcoin & Ethereum addresses. | Open/AKTIF | Open/Confidential | Yes | TBC | Text | TXT | CSV file |
| **Pilot#10 PI** | Financial transactions Data | Financial transactions in operational systems | PI | Confidential | No | Yes | Text/numeric | CSV/ORC | HDFS (batch); REDIS/CASSANDRA (streaming) |
| **Pilot#11, ATOS** | Simulated Urban Mobility Dataset | Simulated Urban mobility data (mainly vehicles CAN Signals) through different scenarios (cities). Captured from SUMO tool) | ATOS | Open | N/A | Yes | Text | JSON | Context Information (Orion CB - NGSI) and Historical data (FIWARE QL) |
| **Pilot#11, ATOS** | CAN Data (Historical Data) | Data collected from vehicle's CAN Bus (80 vehicles driving 4 h/day 1 year). Historical data coming from existing deployments | CTAG | Confidential | Yes | Yes | Text | CSV | MongoDB (suggested) |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#11, ATOS** | Traffic Events (Historical data) | Traffic events published by the city of Vigo and DGT (Historical data related to captured CAN Data) | CTAG | Open | N/A | Yes | Text | JSON | MongoDB (suggested) |
| **Pilot#11, ATOS** | NMEA Data for vehicles (Historical) | Complementary location (GPS, Timestamp, speed, heading...) for Vehicles' CAN Data (Historical data related to captured CAN Data) | CTAG | Confidential | Yes | Yes | Text | CSV | MongoDB (suggested) |
| **Pilot#11, CTAG** | CAN Signals (Live) | CAN data + Driving style info (revolutions, gear, hard braking...)+ Parking (close doors, windows...) + Maintenance | CTAG | Confidential | Yes | Yes | Text | CSV | MongoDB (suggested) |
| **Pilot#11, CTAG** | Traffic Events (Live) | Traffic events published by the city of Vigo and DGT | CTAG | Open | N/A | Yes | Text | JSON | MongoDB (suggested) |
| **PIlot#11, CTAG** | NMEA Data for vehicles (Livel) | Complementary location (GPS, Timestamp, speed, heading...) for Vehicles' CAN Signal | CTAG | Confidential | N/A | Yes | Text | CSV or JSON | MongoDB (suggested) |
| **Pilot#11, DYN** | Motor Insurance Data | Data concerning motor insurance including data from the policies (duration, covers), data from vehicles (licence No, VIN etc.) and data from drivers (age, experience etc.) | DYN | Confidential & Open | Yes | No | Text | CSV | TBD |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#12, iSprint** | Healthentia Live | Measured physical activity (steps, floors, sleep and heart rate) and user reported data from users of Healthentia SaaS who have given consent | iSprint | Confidential | Yes | Models trained for simulator | Text | JSON | MySQL |
| **Pilot#12, iSprint** | Healthentia Simulated | Simulated physical activity and reported data | iSprint | Open | N/A | Already synthetic | Text | JSON | MySQL |
| **Pilot#13, WEA** | SMEWIF | SMEs website information and functionalities | WEA | Confidential | N/A | No | Text | S3 / Dynamo DB | AWS |
| **Pilot#13, WEA** | ROPS | Review and opinions platforms | WEA | Confidential | N/A | No | Text | S3 / Dynamo DB | AWS |
| **Pilot#13, WEA** | EUBD | European SMEs Business Directories | WEA | Confidential | N/A | No | Text | S3 / Dynamo DB | AWS |
| **Pilot#13, WEA** | GIO | SMEs geolocation information and characteristics | WEA | Confidential | N/A | No | Text / Image | S3 / Dynamo DB | AWS |
| **Pilot#13, WEA** | SMSIP | Social media SMEs information and presence | WEA | Confidential | N/A | No | Text | S3 / Dynamo DB | AWS |
| **Pilot#13, WEA** | I&R | Key performance indicators and insurance needs | WEA | Confidential | N/A | No | Text / Image | S3 / Dynamo DB | AWS |
| **Pilot#14, GEN** | Gridded Climate Indices (1/1/1979 to 31/12/2019) | Climate Indices based on the ERA-5 Land and ERA-5 Reanalysis Data | AGRO | Bilateral agreement to issue username/password | N/A | Yes | 3D Gridded Data | NETCDF-CF | THREADDS Data Server |

| INFINITECH Pilot# Dataset Provider | Dataset Name | Dataset (short) description | Owner | License/ Privacy | Anonymized | Capability of Synthetic Data Production | Data Type | Data format | Data store |
|---|---|---|---|---|---|---|---|---|---|
| **Pilot#14, GEN** | EO Data | Earth Observation Data (Sentinel-1,2,3/LandSat-8, MODIS, PROBA-V) for remote damage and crop loss assessment | AGRO | Bilateral agreement to issue username/password | N/A | Yes | Geotiff | GeoJSON | GeoServer |