Data Pitch

H2020-ICT-2016-1

Project number: 732506

D2.1 [DaaS platform]

Coordinator: Stefano Modafferi (University of Southampton IT Innovation Centre)

Quality reviewer: [J. Decis, Dawex]

Deliverable nature:	Other
Dissemination level: (Confidentiality)	Public
Nature	Other
Document URL	
Work package	WP2
Contractual delivery date:	30 Jun 2017
Actual delivery date:	11 Jul 2017
Version:	V1.0
Keywords:	Infrastructure

Table of Contents

ABSTRACT

Different challenges and different scenarios require a high level of degree and flexibility when selecting the appropriate computational and storage solution. Data Pitch supports six different type of infrastructures. We directly control 3 of them while other possibilities include to leave the data in the data provider premises, to store them in the challenge winner's infrastructure or to use a third party commercial service. The solution offered directly by Data Pitch cover different requirements with different level of security, support, and isolation.

EXECUTIVE SUMMARY

This short document is part of the D2.1 deliverable that is classified as “other”. It means that the report complements the installation or better the availability of the infrastructures that are now ready to host data and experiments.

The document describes the different options available and discusses how the data are ingested into the system.

Within the document the acceleration phase is also called experimentation phase and the challenge winners are equivalently called experimenters.

1 INTRODUCTION

This deliverable describes how the data are going to be managed in Data Pitch from an infrastructure point of view. In fact, many options are possible regarding where to store and analyse the data: in the data provider infrastructure, in the experimenter (challenge winner) infrastructure, in the commercial cloud or in the Data Pitch infrastructure. The document briefly presents all of them. DataPitch does not directly control all of the infrastructures. All the Data Pitch infrastructures offer a cloud-like approach where the end-user has control on the assigned resources. They also are able to support the choice and offer the best software stack for the execution of the experiment. There is not a standard proposal and the solution will be identified interacting with the data provider and the challenge winner when the latter is identified.

2 DATA AS A SERVICE (DAAS)

One of the Data Pitch goals is to provide secure, transparent access to a wide range of virtualised data (Data-as-a-Service, or DaaS) and to facilitate a similarly wide range of data analytics. Following the Everything-as-a-Service (XaaS) paradigm, DaaS is based on the idea that data can be provided on-demand to the user regardless of geographic or organizational separation of provider and consumer. With the right DaaS solution, a company can combine data from disparate sources, including their own, and use the resulting knowledge to improve their business.

Access to data is a sensitive topic for any data owner aware of the value hidden within them. It might be the case that the data owners are not yet able to extract the value, but as soon as they understand the potentiality within them, accessing the data become a topic that needs to be carefully addressed.

To support the different needs and requirements that different data providers have Data Pitch offers a wide range of possibilities that are tailored for the specific case:

- Data can be hosted in the data provider infrastructure.
- Data can be hosted in the experimenter infrastructure.
- Data can be hosted in a commercial cloud service.
- Data can be hosted in one of the Data Pitch infrastructures
    - IRIDIS infrastructure
    - IT Innovation infrastructure
    - Secure lab infrastructure

Each of the above solutions addresses different requirements and scenarios and the final decision of where to host the data is taken as a joint decision by the consortium members, the data provider, and the experimenter.

2.1 DATA HOSTED IN THE DATA PROVIDER INFRASTRUCTURE

This case is likely to be the most common, as it gives the full control of the data access to the data owner. Data Pitch will have a supporting role if required by the data owner. Though this seems an ideal situation, some consideration needs also to be given to the fact that solutions implemented in the acceleration phase are meant to be highly innovative and possibly business and technology disruptive. This approach requires the creation of sandboxed parts of the infrastructure where the experiment can happen. So the price a data provider will pay for retaining the full control can include an increment of the computational power and possibly of the storage space. To mitigate this problem it is always possible to use the Southampton infrastructure as discussed in the next paragraphs.

2.2 DATA HOSTED IN THE DATA EXPERIMENTER INFRASTRUCTURE

Data may be provided directly to the experimenters. This is likely to happen when the experiment, possibly following instructions from the data provider, does not require the access to the full spectrum of data. For instance the challenge can impact a subset of the data possibly less relevant from a sensitivity point of view, or it can be around pseudonymized data or it can contains only historical information. In all these cases it can be convenient that the data are stored directly within the experimenter facility. Potential issues include problems of scalability and the exclusion of experimenters who do not have enough hosting or computational capacity, but we should not exclude scenarios in which this is a good option. The options of Data Pitch hosting the data is always available if the data providers wants to have the experiment outside their facility and the experimenter has not enough capacity to host it. DataPitch role in this scenario is supporting the data transfer between provider and experimenter and check that all the legal requirements of data hosting are fulfilled.

2.3 Data hosted in the data Commercial Cloud service

This case is listed as one of the possibility and can be the result of two different scenarios. The data owners can already use a commercial cloud to host the data. In this case the situation is very similar to the one above where the infrastructure is managed directly by the data owner. The role of Data Pitch will be to support the data owner as required. A slightly different and less likely case happens when it is a common decision of data owner and experimenter to make use of a commercial cloud rather than an existing infrastructure or the one offered by Data Pitch. It is hard to think of a real possible case when this latter situation will be the agreed solution. Data Pitch is anyway prepared to support it. The payment of the service needs to be covered either by the data owner or by the grant issued to the experimenter.

2.4 Data hosted in the data Pitch infrastructure

This case is split into three possible different solutions offering an increasing level of security. Anyway, security will not be the only parameter considered when choosing the infrastructure. In fact, the availability of resources and computational power is higher in IRIDIS, while other aspects like the risk of being disruptive and isolation as well as the possibility of a dedicated support are better addressed in the IT Innovation infrastructure and the top of the security with physical control access is implemented in the Secure Lab. All the infrastructures follows the data as a service paradigm offering a cloud-like approach that allows the end-user to manage their (virtual) machines and their own data. All the infrastructure supports the typical big-data software stack and are available for customization and specific request that each single experiment might have.

2.4.1 IRIDIS infrastructure

The University's IRIDIS Compute Cluster provides High Performance Computing facilities in a professionally managed service environment. IRIDIS currently has the following computational power: i) 750 x 2.6 GHz 16-core nodes with ~62 GB usable memory; 12 x GPU nodes (2.6 GHz 16-core nodes with ~62 GB usable memory + two K20 GPU cards); 12 x Phi nodes (2.6 GHz 16-core nodes with ~62 GB usable memory + phi cards); 4 x "fat" nodes with 32 cores and ~250GB usable memory.

The share of the computational power accessible by the Data Pitch experiments depends on the load of the infrastructure at the time of the experiment. IRIDIS is constantly investing in new hardware and advanced level of support. An incremental process of the computational and storing capacity is already in place and eventually (planned Jan 2018) the total figures will be increased by four times.

The engagement with the infrastructure will be done through the WP2 work package that have already in place all the agreement for the start of experiment also at short notice.

2.4.2 IT Innovation infrastructure

The IT Innovation infrastructure actually sits beyond a further proxy layer with respect to IRIDIS. This infrastructure offers an increased level of isolation, the possibility of more disruptive experiment and a dedicated support.

IT Innovation infrastructure currently offers a capacity of ~80 cores with > 320 GByte RAM and > 20 TByte Hard disk. This is an expandable cluster and it offers a great level of flexibility in terms of the software stack, as it is possible to create ad-hoc solutions if an experiment requires this characteristic.

The IT Innovation centre is part of the DataPitch consortium and the engagement for the use of the infrastructure is straightforward as It Innovation is the leader of Wp2.

2.4.3 Secure lab infrastructure

For cases where the commercial sensitivity and/or specific accreditation (e.g. for health clinic trials) the Secure Lab infrastructure is the right choice. The infrastructure requires physical authentication either in Southampton or in an authorized location and it is BIL3 and ISO27001–compliant Tier2 data centre with

robust security measures to cater for data up to and including UK “OFFICIAL-SENSITIVE” or the equivalent. The infrastructure has also recently obtained the UK Information Governance Toolkit.

With robust security measures (CCTV, 24x7 onsite security, rigorous access controls), the laboratory provides secure scalable enterprise class storage, with asynchronous replication of data between the primary and Disaster Recovery (DR) sites. The current storage capacity is ~70 TB.

The engagement with the infrastructure will be done through the WP2 work package that have already in place all the agreement for the start of experiment also at short notice.

2.5 Data Ingestion

Different possibilities exist for supporting the data ingestion. This is necessary when an infrastructure different from the data provider one is going to be used to implement the experiment. The ingestion can be an off-line ingestion with or without an update frequency, and a real time ingestion.

2.5.1 Off-line ingestion

In this case, a dataset is provided by the data provider to the infrastructure owner. It can be a physical drive if the size is too big or the dataset can be transferred as a set of file through an ftp server and then conveniently stored in the target infrastructure. In Data Pitch the off-line ingestion currently happens though an ad-hoc process to support the maximum flexibility. Also the update frequency depends on the type of experiment. If the dataset needs to be updated during the experiment, the process will be automated with a scheduled task to download and ingest the new data when agreed. An automatic process will also be implemented if a number of standard cases is identified when discussing with the data providers. So to reduce the effort, leaving the ad-hoc process for a limited number of cases.

2.5.2 real time ingestion

One of the most challenging and also interesting case if the execution of analytic functions on data streams. Data Pitch will support the automatic download of data stream leveraging on the corresponding ability provided by the DAWEX platform. In fact, the platform works as a bridge that directs a stream of data produced by a source to a selected target. In the case of Data Pitch the source will be the data provider and the destination will be the target infrastructure where the experiment is implemented.

2.6 data sampling

Data Pitch offers the Data providers to describe their data within the Data Pitch Catalogue (cf. Wp3). On top of this possibility a provider can also give Data Pitch a sample of their data to be made available for the download during the time when a challenge is open. The possibility can be activated at any time, though currently no data providers have expressed their willing to share a sample of their data

Data Pitch current supports this possibility by using the It Innovation infrastructure, while for the second round of the challenges the Dawex platform is going to host both the catalogue and the possibility of data sampling downloads.

3 CONCLUSION

This document discussed the different solutions available in Data Pitch to host data and experiments. The variety of scenarios and requirements that the challenges present require a flexible approach and an interaction with all the involved parties to identify the best solution. For this reason, several steps are implemented in an ad-hoc fashion and large part of the information will be available when the winner will be selected and the type and goal of the experiment will be clearly defined. The infrastructures directly controlled by Data Pitch are now available and ready to engage when the actual acceleration period will start.