Enabling a data-informed public sector:

*From hype to action using the **Big Data Test Infrastructure (BDTI)***

Maria Claudia BODINO, BDTI project officer – European Commission
mariaclaudia.bodino@ec.europa.eu

European Commission

Business Owner:
DG CNECT
Directorate-General for Communications Networks, Content and Technology

Service Provider:
DG DIGIT
Directorate-General for Digital Services

Road Map

1 Policy context

2 BDTI in a nutshell
- Its context and why use it

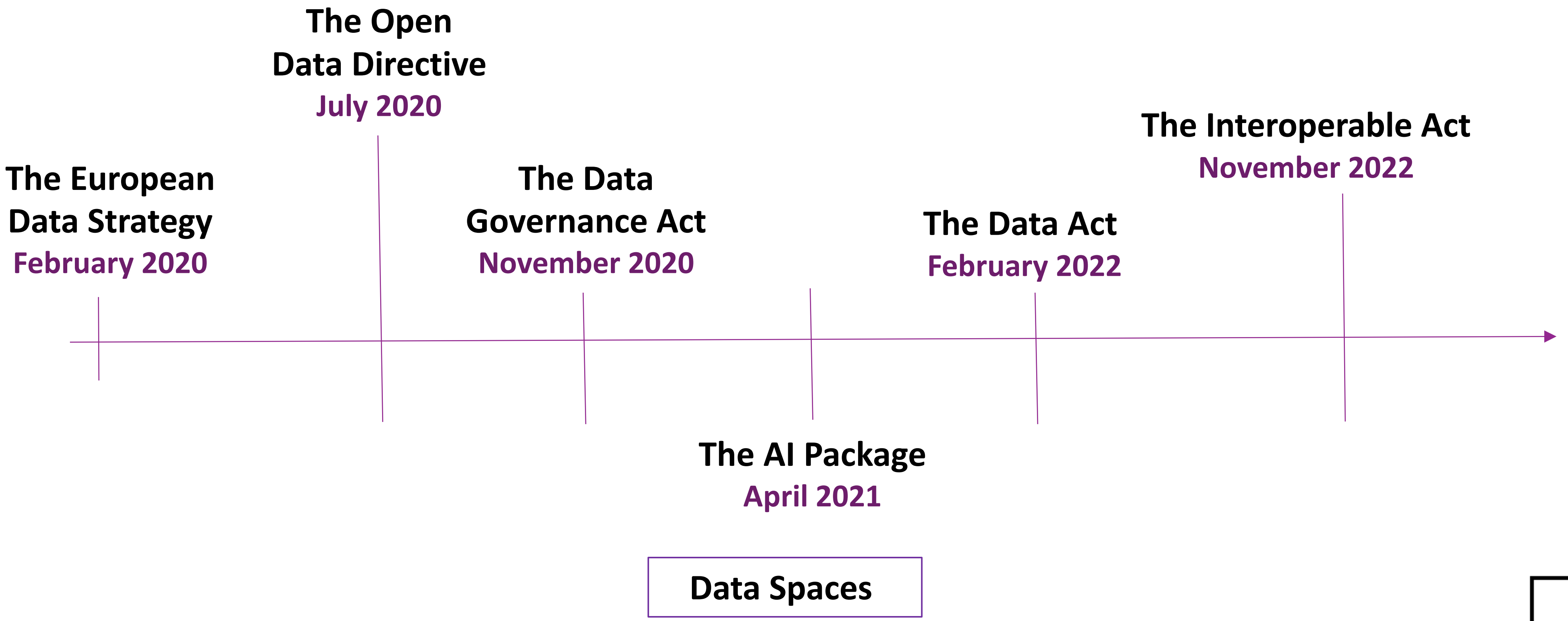
3 BDTI in practice
- Access and overview of the BDTI portal
- Concrete application of the BDTI

4 BDTI's community
- Developing the BDTI community and how can you help us

1 Policy context

1

Policy timeline



The European Data Strategy
February 2020

The Open Data Directive
July 2020

The Data Governance Act
November 2020

The AI Package
April 2021

The Data Act
February 2022

The Interoperable Act
November 2022

Data Spaces

Policy vision

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

Big Data Test Infrastructure (BDTI) in a nutshell: its context

The BDTI is funded by the the Digital Europe Program (DEP), an **EU funding programme** (€7.5 bn) focused on **bringing digital technology** to businesses, citizens and **public administrations**.

The DEP provides strategic funding in **five crucial areas**:

- **High performance computing**
- **Cybersecurity**
- **Artificial intelligence**
  (Cloud, data and AI)
- **Advanced digital skills**
- **Deployment and wide use of digital technologies**

Learn more about the Digital Europe Programme: https://digital-strategy.ec.europa.eu/en/activities/digital-programme

2

BDTI in a nutshell

- Its context and why use it

Public Sector Information and the role of Data analytics

Data is **everywhere** and growing at an unprecedent pace.

- **Big Data**: 3V - **V**olume, **V**ariety, **V**elocity

Data is a key ingredient for **services, products, and effective policy making.**

There is an ambition to create a **single European market for data** and make more data available through powerful and trustworthy infrastructures and technologies, **in line with EU values and regulations, to support citizens, public sector and companies**.

Learn more about the Digital Europe Programme: https://digital-strategy.ec.europa.eu/en/activities/digital-programme

Public Sector Information and the role of Data analytics

Big Data is identified as

1. Data created by **private citizens** in their interaction
2. Data collected by **sensors and automatically** transmitted online
3. Data collected by **public bodies** during their operation

(Mergel et al., 2016)

In **public policy** Big Data is associated with

1. new **formats**
2. **quality**
3. **availability**

of administrative data

(Pirog, 2014).

Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, *76*(6), 928-937.

Pirog, M. A. (2014). Data will drive innovation in public policy and management research in the next decade. *Journal of Policy Analysis and Management*, 537-543.

What is the Big Data Test Infrastructure (BDTI) ?

Not **only** for big data, for **public sector in general (i.e. open data)**

Six months free of charge service for EU public administrations *

Ready-to-use **data analytics stack** and support

Cloud platform based on **open-source** tools

To help the public sector **to derive insights from data** and accelerate transition towards **data- informed decision making.**

* The cost of the pilot project must fit within the funding boundaries of the BDTI pilot budget

Big Data Test Infrastructure Objectives

Objectives

- Increase the easy accessibility, interoperability, quality and usability of public sector information in compliance with the requirement of the **Open Data Directive**
- Boost the **re-use and combination of open public data** across the EU for the development of information products and services, including AI applications
- High Value Datasets – Open Data Directive
- Testing **Business-to-Government** (B2G) data sharing collaborations for the **public good**
- Data Space Support Centre: explore and experiment with your data*
  - BDTI provides a safe **testing environment to run big data experiment**s for data space customers

* https://joinup.ec.europa.eu/collection/semic-support-centre/data-spaces

2

Who is the Big Data Test Infrastructure (BDTI) for?

European Public Administrations

All European Public Administrations at local, regional and national level can independently apply for a BDTI pilot project

Ecosystem with academia and private sector

Academia, spin-off, startups can apply for pilot projects as long as there there is a **clear collaboration** with a Public Administration which will be the main point of contact for the project (Master/PhD, GovTech startups)

Are you working for a public administration in need of infrastructure for data analytics?

Contact us:

EC-BDTI-PILOTS@ec.europa.eu

2

Why use the BDTI ?

Benefit of six months free of charge service, including **advisory and technical** support during the duration of the pilot

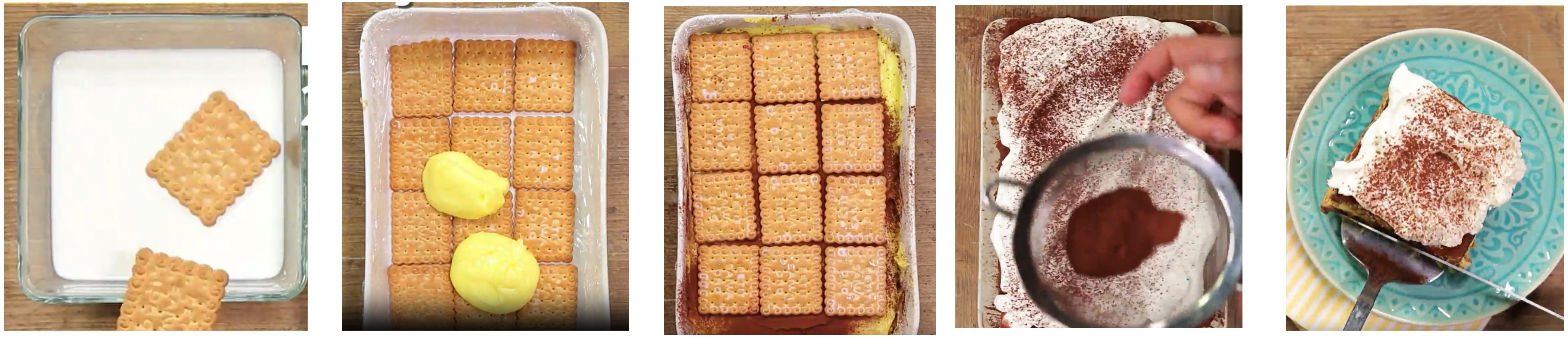
Experiment with data analytics using high **performance infrastructure** that leverages the power of the **elastic cloud**

Receive guidance to move from a pilot to a **production-ready** process – **EXIT package**

Test your idea → Extract value → Create knowledge

Why use the BDTI ?

Data → Information → Presentation → Knowledge

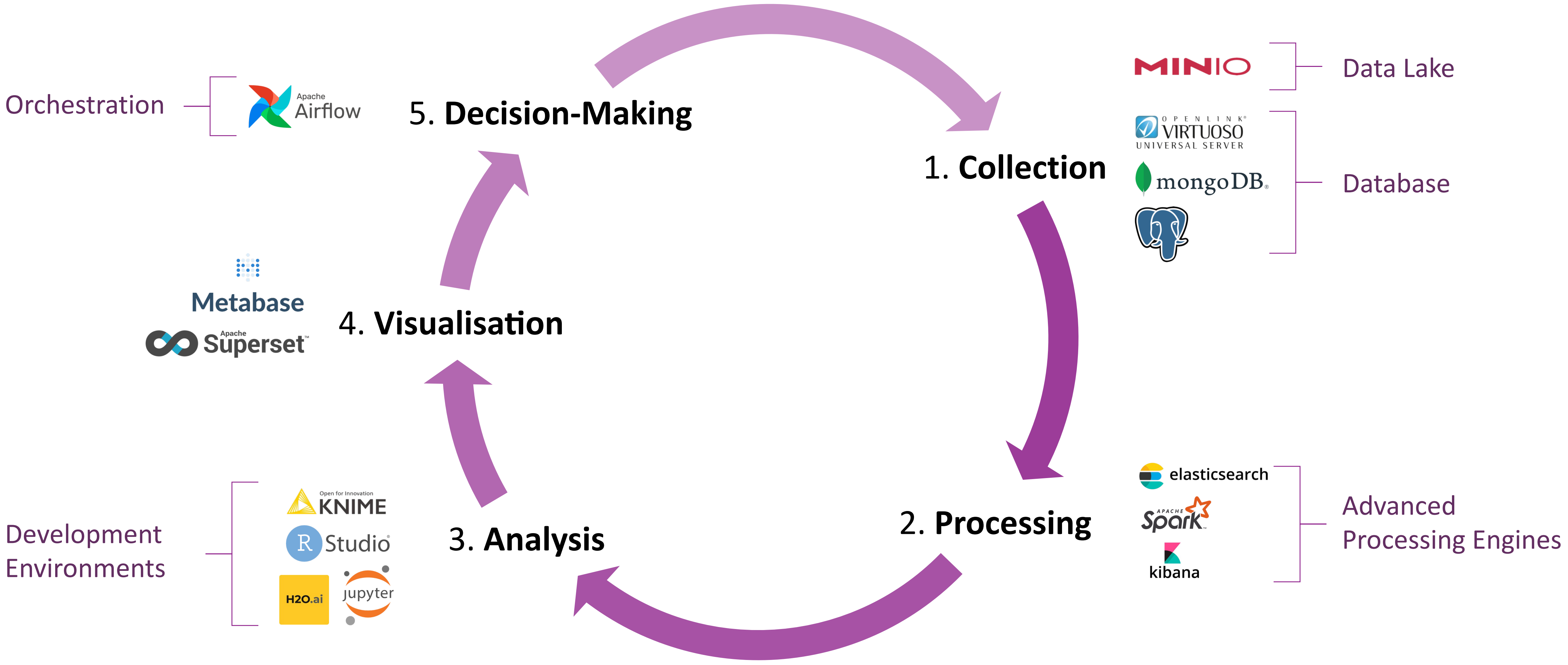


You have the key ingredients (datasets),
we provide you the best tool to generate amazing recipes.

https://funtip.giallozafferano.it/Torta-mattonella.html

2

With its open source tools, BDTI supports you throughout your data journey



1. Collection

2. Processing

3. Analysis

4. Visualisation

5. Decision-Making

Data Lake

Database

Advanced Processing Engines

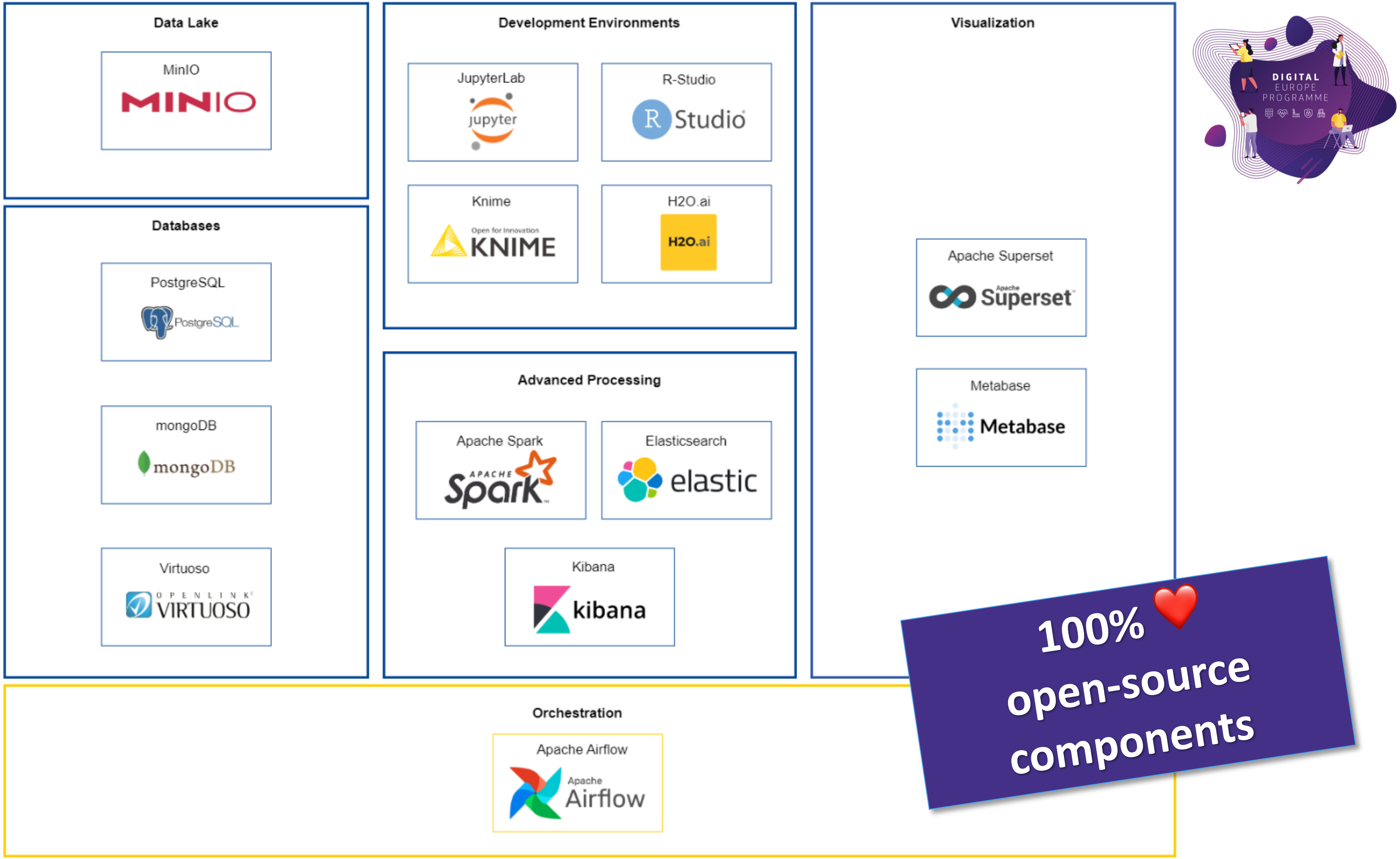
Development Environments

Orchestration

2

BDTI's Data Analytics Stack

DIGITAL EUROPE PROGRAMME



Data Lake

- MinIO

Databases

- PostgreSQL
- mongoDB
- Virtuoso

Development Environments

- JupyterLab
- R-Studio
- Knime
- H2O.ai

Advanced Processing

- Apache Spark
- Elasticsearch
- Kibana

Visualization

- Apache Superset
- Metabase

Orchestration

- Apache Airflow

100% ❤️ open-source components

https://big-data-test-infrastructure.ec.europa.eu/service-offering_en

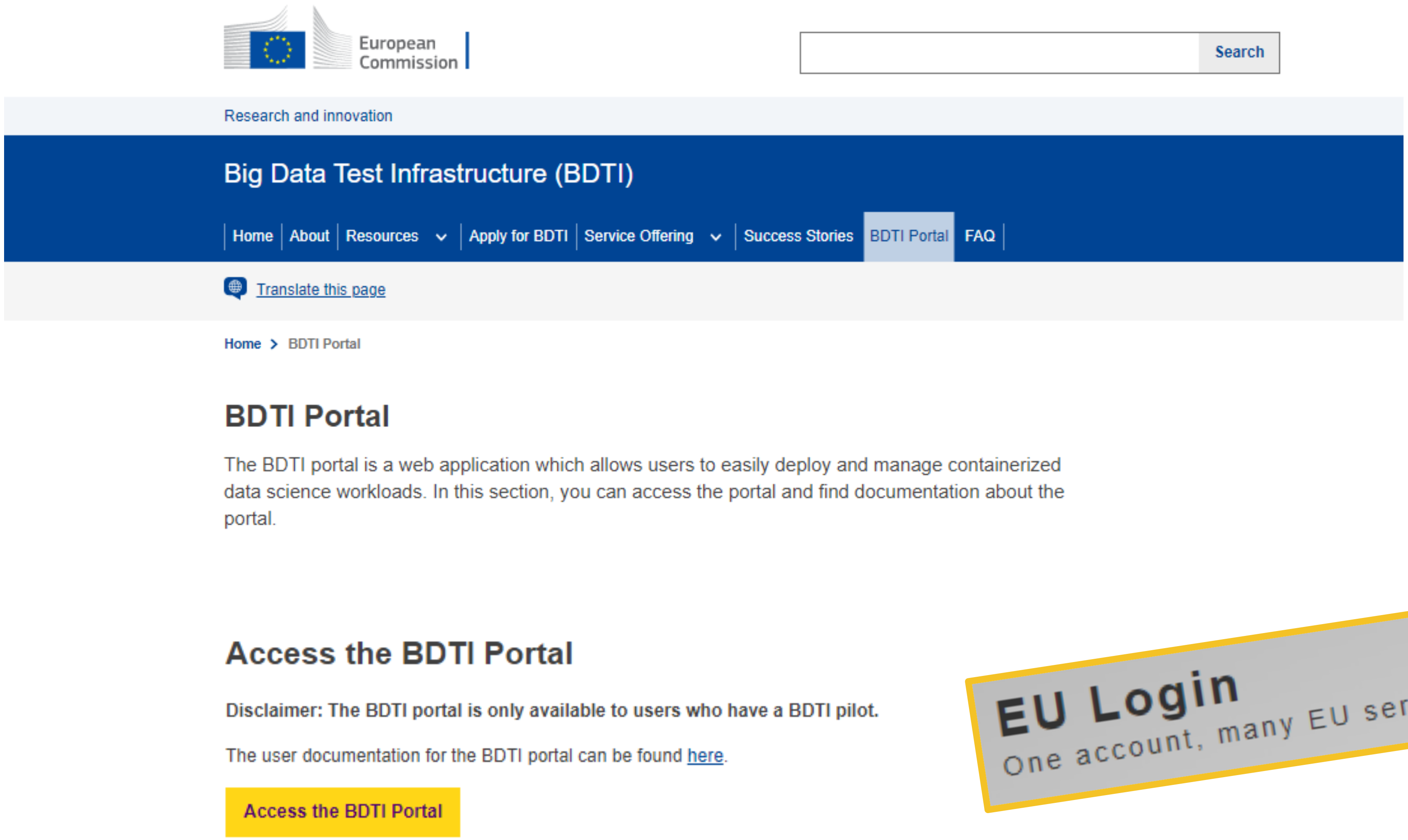
3

BDTI in practice

- Access and overview of the BDTI portal
- Concrete application of the BDTI

3

Access to BDTI portal directly from your browser (EU Login integration)



Big Data Test Infrastructure (BDTI)

BDTI Portal

The BDTI portal is a web application which allows users to easily deploy and manage containerized data science workloads. In this section, you can access the portal and find documentation about the portal.

Access the BDTI Portal

Disclaimer: The BDTI portal is only available to users who have a BDTI pilot.

The user documentation for the BDTI portal can be found here.

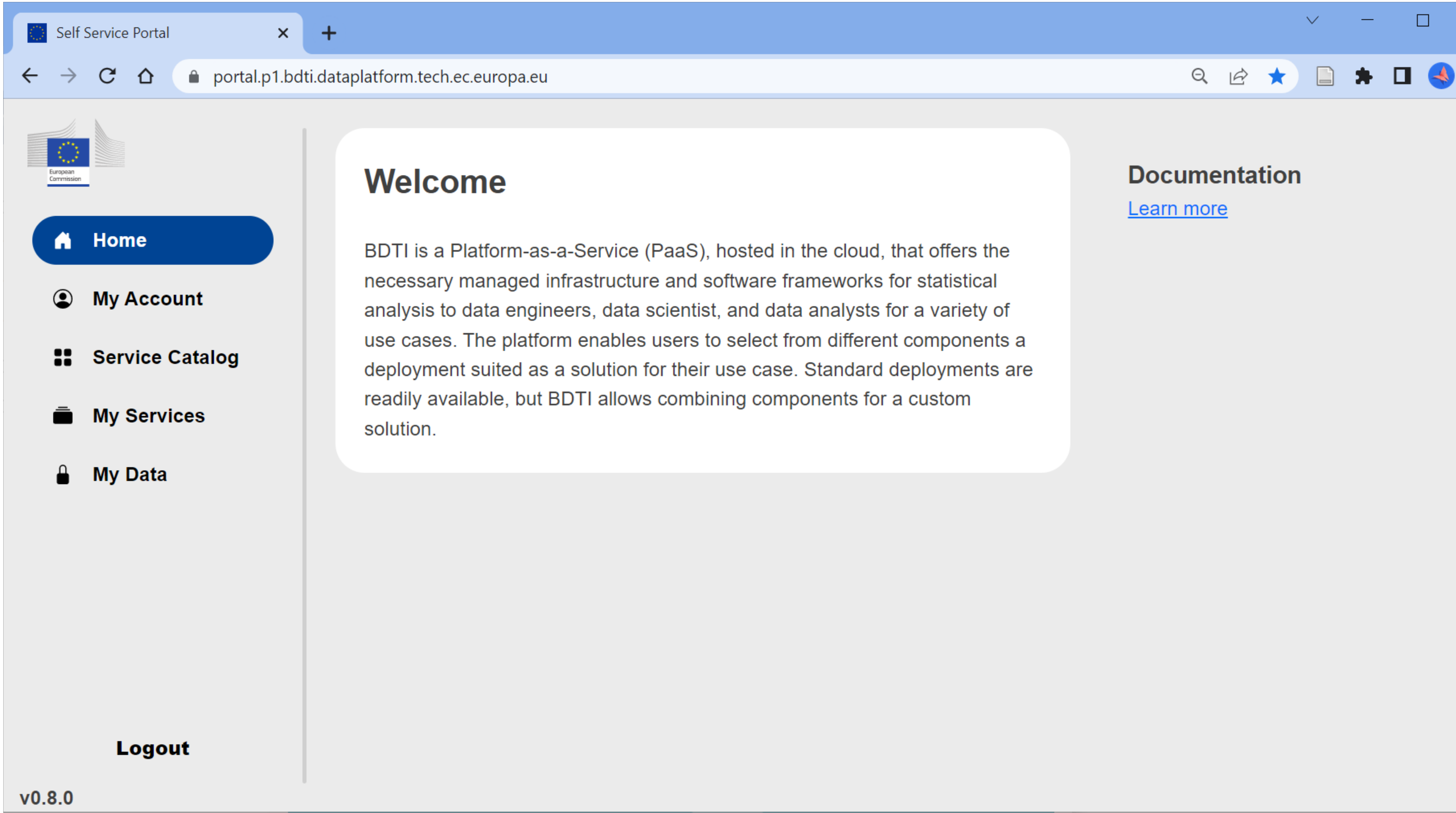
Access the BDTI Portal

EU Login
One account, many EU services

For teams part of BDTI pilots

3

The BDTI portal



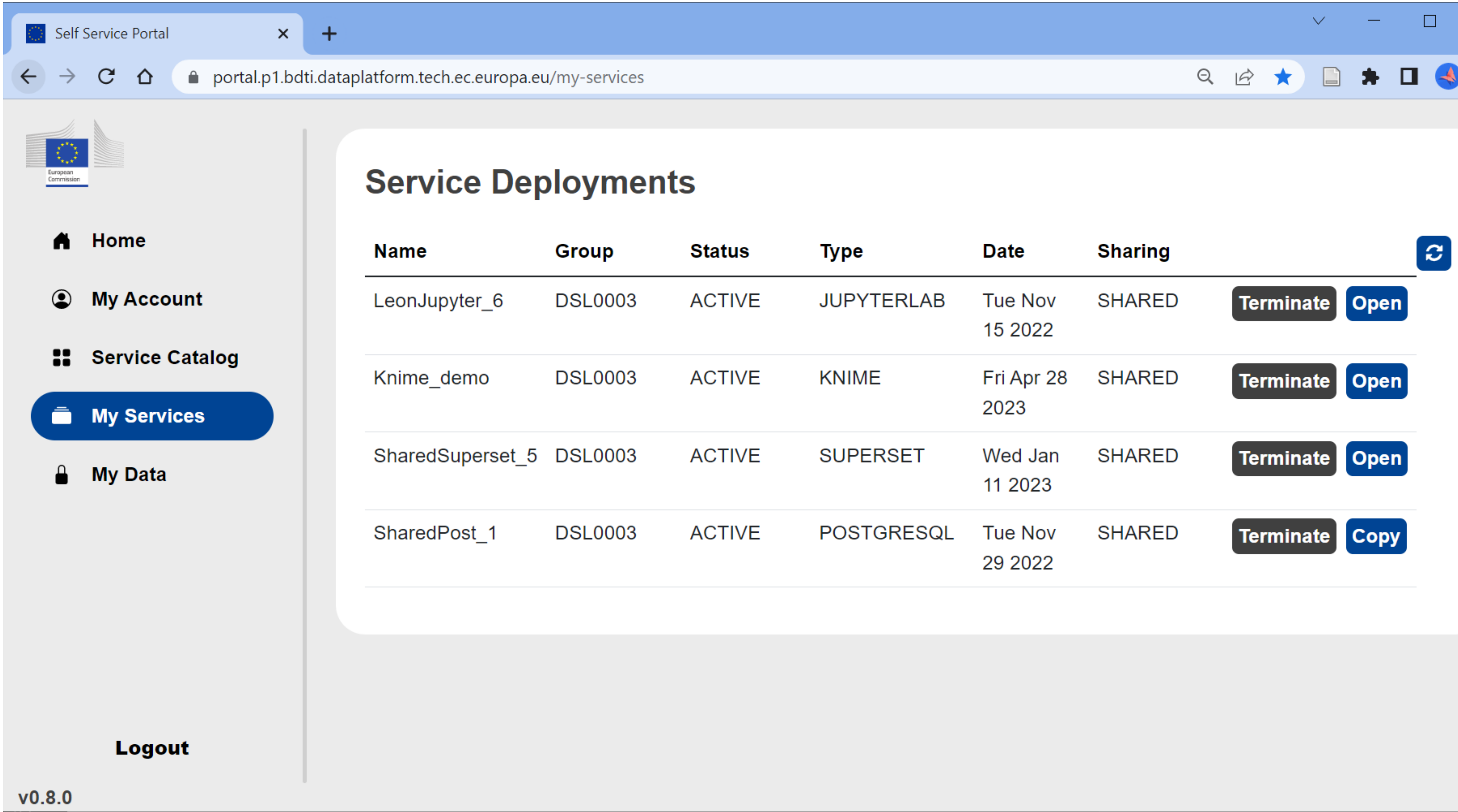
Welcome

BDTI is a Platform-as-a-Service (PaaS), hosted in the cloud, that offers the necessary managed infrastructure and software frameworks for statistical analysis to data engineers, data scientist, and data analysts for a variety of use cases. The platform enables users to select from different components a deployment suited as a solution for their use case. Standard deployments are readily available, but BDTI allows combining components for a custom solution.

Documentation

Learn more

The BDTI portal: My Services

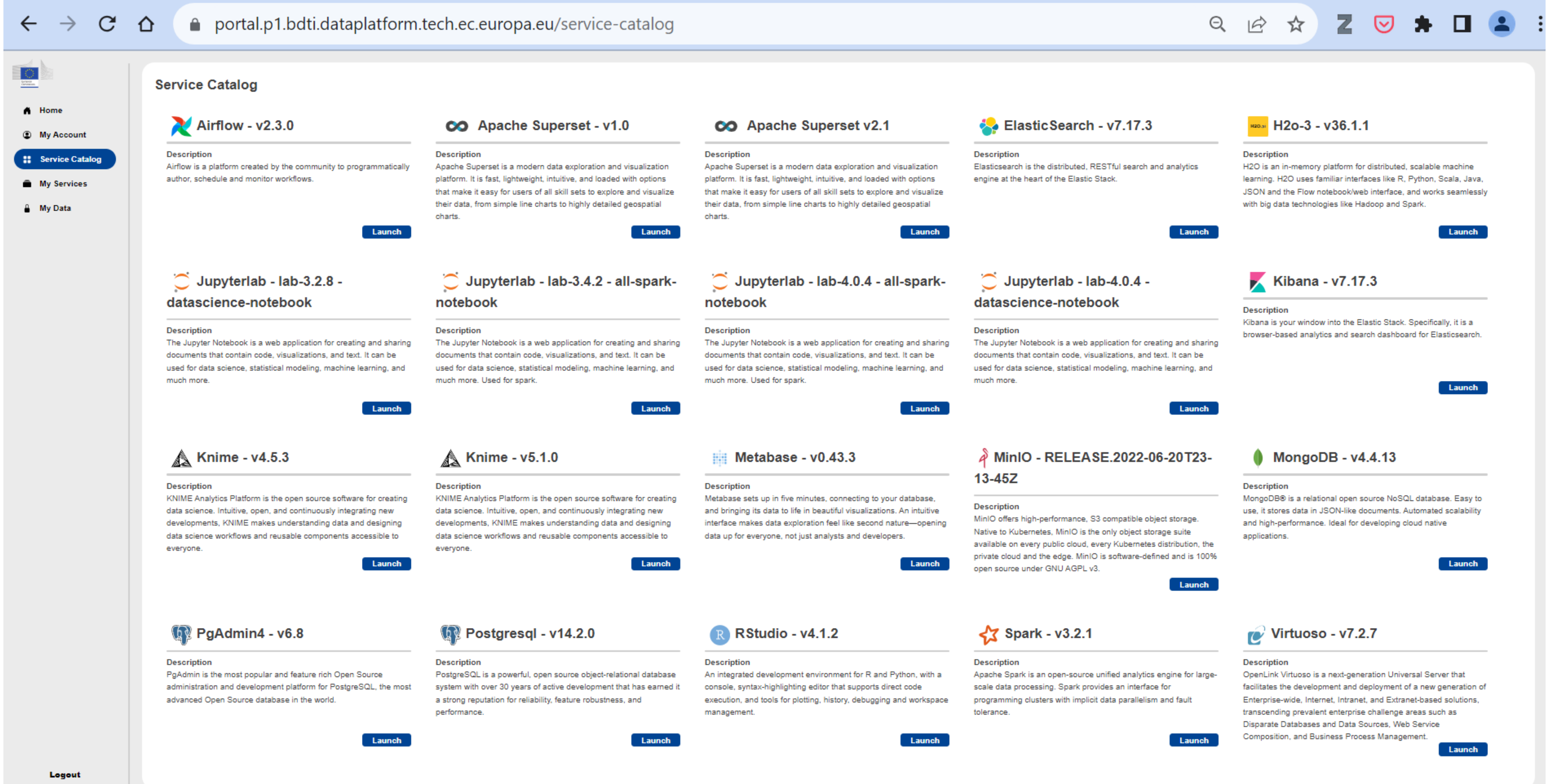


Service Deployments

Name	Group	Status	Type	Date	Sharing	
LeonJupyter_6	DSL0003	ACTIVE	JUPYTERLAB	Tue Nov 15 2022	SHARED	Terminate Open
Knime_demo	DSL0003	ACTIVE	KNIME	Fri Apr 28 2023	SHARED	Terminate Open
SharedSuperset_5	DSL0003	ACTIVE	SUPERSET	Wed Jan 11 2023	SHARED	Terminate Open
SharedPost_1	DSL0003	ACTIVE	POSTGRESQL	Tue Nov 29 2022	SHARED	Terminate Copy

3

The BDTI portal: service catalogue


portal.p1.bdti.dataplatform.tech.ec.europa.eu/service-catalog
Home
My Account
Service Catalog
My Services
My Data
Logout

Service Catalog

Airflow - v2.3.0

Description
Airflow is a platform created by the community to programmatically author, schedule and monitor workflows.

Launch

Apache Superset - v1.0

Description
Apache Superset is a modern data exploration and visualization platform. It is fast, lightweight, intuitive, and loaded with options that make it easy for users of all skill sets to explore and visualize their data, from simple line charts to highly detailed geospatial charts.

Launch

Apache Superset v2.1

Description
Apache Superset is a modern data exploration and visualization platform. It is fast, lightweight, intuitive, and loaded with options that make it easy for users of all skill sets to explore and visualize their data, from simple line charts to highly detailed geospatial charts.

Launch

ElasticSearch - v7.17.3

Description
Elasticsearch is the distributed, RESTful search and analytics engine at the heart of the Elastic Stack.

Launch

H2o-3 - v36.1.1

Description
H2O is an in-memory platform for distributed, scalable machine learning. H2O uses familiar interfaces like R, Python, Scala, Java, JSON and the Flow notebook/web interface, and works seamlessly with big data technologies like Hadoop and Spark.

Launch

Jupyterlab - lab-3.2.8 - datascience-notebook

Description
The Jupyter Notebook is a web application for creating and sharing documents that contain code, visualizations, and text. It can be used for data science, statistical modeling, machine learning, and much more.

Launch

Jupyterlab - lab-3.4.2 - all-spark-notebook

Description
The Jupyter Notebook is a web application for creating and sharing documents that contain code, visualizations, and text. It can be used for data science, statistical modeling, machine learning, and much more. Used for spark.

Launch

Jupyterlab - lab-4.0.4 - all-spark-notebook

Description
The Jupyter Notebook is a web application for creating and sharing documents that contain code, visualizations, and text. It can be used for data science, statistical modeling, machine learning, and much more. Used for spark.

Launch

Jupyterlab - lab-4.0.4 - datascience-notebook

Description
The Jupyter Notebook is a web application for creating and sharing documents that contain code, visualizations, and text. It can be used for data science, statistical modeling, machine learning, and much more.

Launch

Kibana - v7.17.3

Description
Kibana is your window into the Elastic Stack. Specifically, it is a browser-based analytics and search dashboard for Elasticsearch.

Launch

Knime - v4.5.3

Description
KNIME Analytics Platform is the open source software for creating data science. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone.

Launch

Knime - v5.1.0

Description
KNIME Analytics Platform is the open source software for creating data science. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components accessible to everyone.

Launch

Metabase - v0.43.3

Description
Metabase sets up in five minutes, connecting to your database, and bringing its data to life in beautiful visualizations. An intuitive interface makes data exploration feel like second nature—opening data up for everyone, not just analysts and developers.

Launch

MinIO - RELEASE.2022-06-20T23-13-45Z

Description
MinIO offers high-performance, S3 compatible object storage. Native to Kubernetes, MinIO is the only object storage suite available on every public cloud, every Kubernetes distribution, the private cloud and the edge. MinIO is software-defined and is 100% open source under GNU AGPL v3.

Launch

MongoDB - v4.4.13

Description
MongoDB® is a relational open source NoSQL database. Easy to use, it stores data in JSON-like documents. Automated scalability and high-performance. Ideal for developing cloud native applications.

Launch

PgAdmin4 - v6.8

Description
PgAdmin is the most popular and feature rich Open Source administration and development platform for PostgreSQL, the most advanced Open Source database in the world.

Launch

Postgresql - v14.2.0

Description
PostgreSQL is a powerful, open source object-relational database system with over 30 years of active development that has earned it a strong reputation for reliability, feature robustness, and performance.

Launch

RStudio - v4.1.2

Description
An integrated development environment for R and Python, with a console, syntax-highlighting editor that supports direct code execution, and tools for plotting, history, debugging and workspace management.

Launch

Spark - v3.2.1

Description
Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance.

Launch

Virtuoso - v7.2.7

Description
OpenLink Virtuoso is a next-generation Universal Server that facilitates the development and deployment of a new generation of Enterprise-wide, Internet, Intranet, and Extranet-based solutions, transcending prevalent enterprise challenge areas such as Disparate Databases and Data Sources, Web Service Composition, and Business Process Management.

Launch

3

BDTI Demonstrator:
Towards a data-Informed Government Spending

Goal:

Show how the BDTI can be used by different users (at different levels of complexity) to **derive insights from government spendings to take data-informed actions**

A user-centered approach:

- Elena and Daniel, public servants
- Low data literacy skills
- **Problem**: high government spending in public lighting
- **Solution**: how to optimize public lighting to reduce government spending

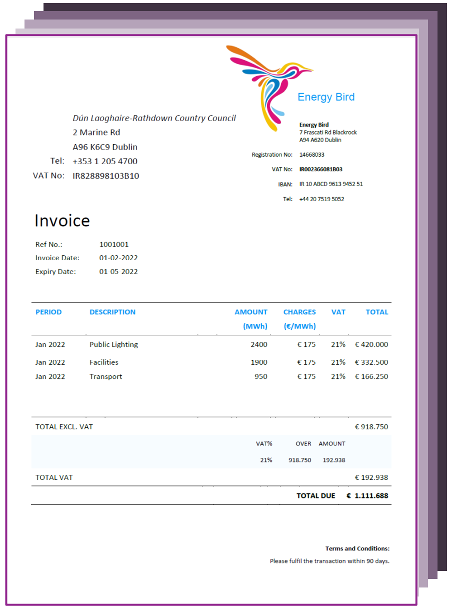
3

Open for Innovation
KNIME

ETL - Data extraction from non-machine readable PDF files

PostgreSQL

Storage & structuring of collected data



Energy Bird

Dún Laoghaire-Rathdown Country Council
2 Marine Rd
A96 K6C9 Dublin
Tel: +353 1 205 4700
VAT No: IR828898103B10

Energy Bird
7 Frascati Rd Blackrock
A94 A620 Dublin

Registration No: 14668033
VAT No: **IR002366081B03**
IBAN: IR 10 ABCD 9613 9452 51
Tel: +44 20 7519 5052

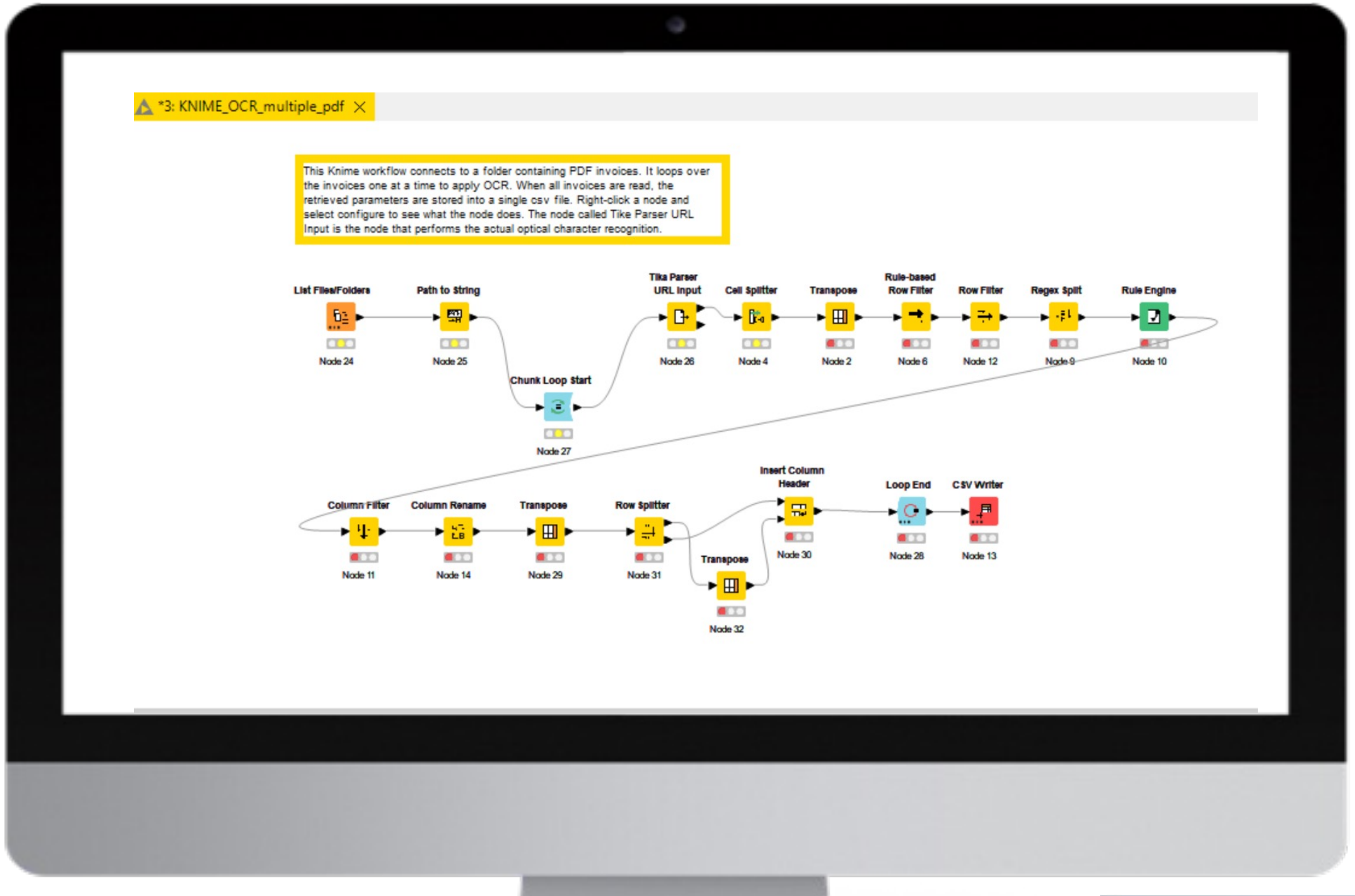
Invoice

Ref No.: 1001001
Invoice Date: 01-02-2022
Expiry Date: 01-05-2022

PERIOD	DESCRIPTION	AMOUNT (MWh)	CHARGES (€/MWh)	VAT	TOTAL
Jan 2022	Public Lighting	2400	€ 175	21%	€ 420.000
Jan 2022	Facilities	1900	€ 175	21%	€ 332.500
Jan 2022	Transport	950	€ 175	21%	€ 166.250

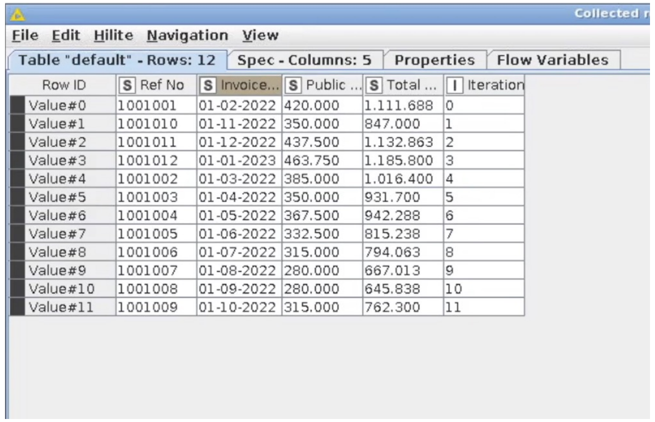
TOTAL EXCL. VAT			€ 918.750
VAT%	OVER	AMOUNT	
21%	918.750	192.938	
TOTAL VAT			€ 192.938
		TOTAL DUE	**€ 1.111.688**

Terms and Conditions:
Please fulfil the transaction within 90 days.

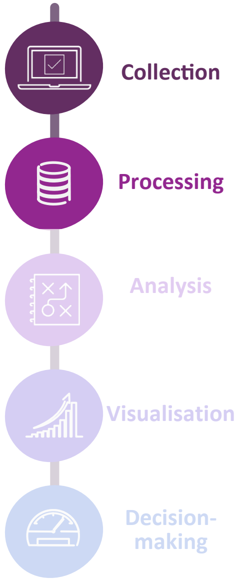


*3: KNIME_OCR_multiple_pdf

This Knime workflow connects to a folder containing PDF invoices. It loops over the invoices one at a time to apply OCR. When all invoices are read, the retrieved parameters are stored into a single csv file. Right-click a node and select configure to see what the node does. The node called Tike Parser URL Input is the node that performs the actual optical character recognition.



Row ID	S Ref No	S Invoice...	S Public ...	S Total ...	I Iteration
Value#0	1001001	01-02-2022	420.000	1.111.688	0
Value#1	1001010	01-11-2022	350.000	847.000	1
Value#2	1001011	01-12-2022	437.500	1.132.863	2
Value#3	1001012	01-01-2023	463.750	1.185.800	3
Value#4	1001002	01-03-2022	385.000	1.016.400	4
Value#5	1001003	01-04-2022	350.000	931.700	5
Value#6	1001004	01-05-2022	367.500	942.288	6
Value#7	1001005	01-06-2022	332.500	815.238	7
Value#8	1001006	01-07-2022	315.000	794.063	8
Value#9	1001007	01-08-2022	280.000	667.013	9
Value#10	1001008	01-09-2022	280.000	645.838	10
Value#11	1001009	01-10-2022	315.000	762.300	11



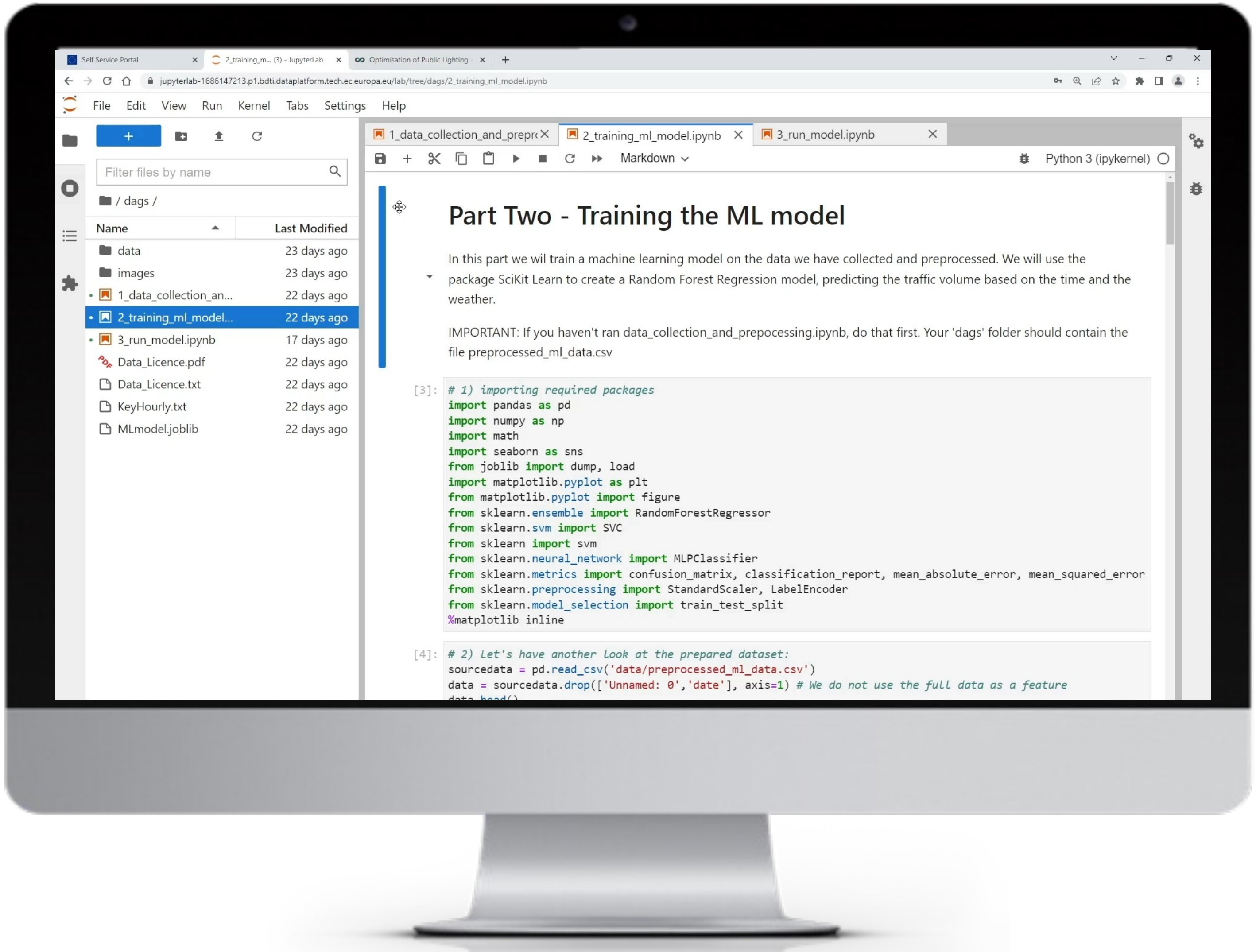
3

jupyter

To train Machine Learning (ML) models to analyse big (or small) amount of data

PostgreSQL

Stores the newly analysed data



Self Service Portal
2_training_m... (3) - JupyterLab
Optimisation of Public Lighting -
jupyterlab-1686147213.p1.bdti.dataplatform.tech.ec.europa.eu/lab/tree/dags/2_training_ml_model.ipynb
File
Edit
View
Run
Kernel
Tabs
Settings
Help
+
Filter files by name
/ dags /
1_data_collection_and_prepr
2_training_ml_model.ipynb
3_run_model.ipynb
Markdown
Python 3 (ipykernel)

Name	Last Modified
data	23 days ago
images	23 days ago
1_data_collection_an...	22 days ago
2_training_ml_model...	22 days ago
3_run_model.ipynb	17 days ago
Data_Licence.pdf	22 days ago
Data_Licence.txt	22 days ago
KeyHourly.txt	22 days ago
MLmodel.joblib	22 days ago

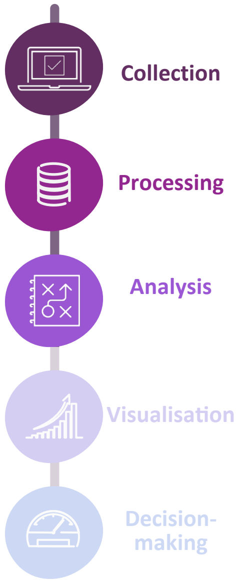
Part Two - Training the ML model

In this part we wil train a machine learning model on the data we have collected and preprocessed. We will use the package SciKit Learn to create a Random Forest Regression model, predicting the traffic volume based on the time and the weather.

IMPORTANT: If you haven't ran data_collection_and_prepocessing.ipynb, do that first. Your 'dags' folder should contain the file preprocessed_ml_data.csv

```
[3]: # 1) importing required packages
import pandas as pd
import numpy as np
import math
import seaborn as sns
from joblib import dump, load
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVC
from sklearn import svm
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix, classification_report, mean_absolute_error, mean_squared_error
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.model_selection import train_test_split
%matplotlib inline
```

```
[4]: # 2) Let's have another look at the prepared dataset:
sourcedata = pd.read_csv('data/preprocessed_ml_data.csv')
data = sourcedata.drop(['Unnamed: 0','date'], axis=1) # We do not use the full data as a feature
```



Collection

Processing

Analysis

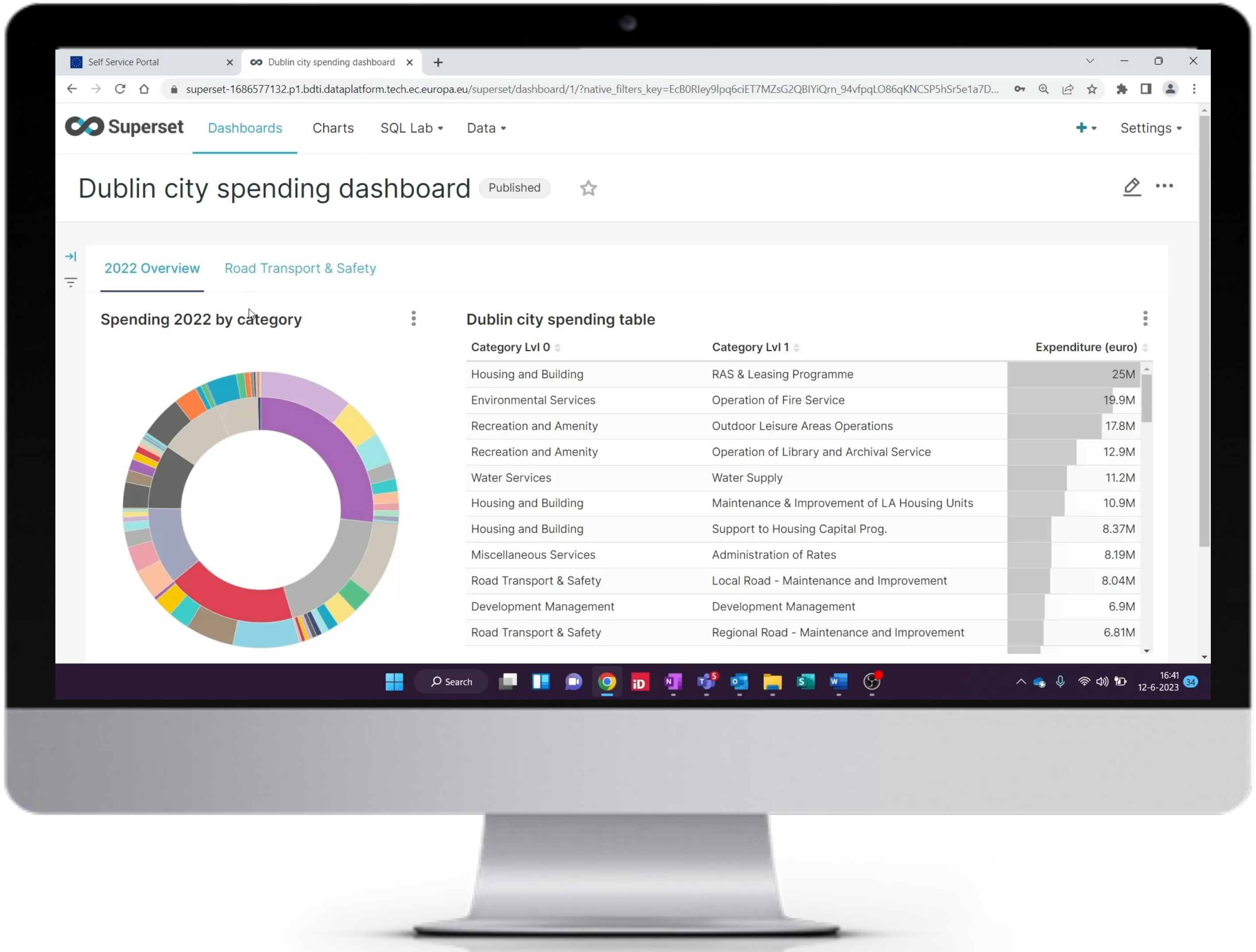
Visualisation

Decision-making

3

Apache Superset™

Interactive overview of expenditures per category
Ability to benchmark against different data sets


Self Service Portal
Dublin city spending dashboard
superset-1686577132.p1.bdti.dataplatform.tech.ec.europa.eu/superset/dashboard/1/?native_filters_key=EcB0Rley9lpq6ciET7MZsG2QBIYiQrn_94vfpqLO86qKNCSP5hSr5e1a7D...
Superset
Dashboards
Charts
SQL Lab
Data
Settings
Published
Search
16:41
12-6-2023

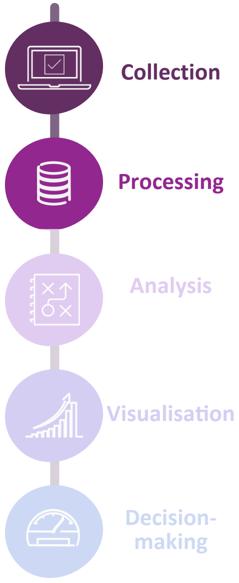
Dublin city spending dashboard

2022 Overview | Road Transport & Safety

Spending 2022 by category

Dublin city spending table

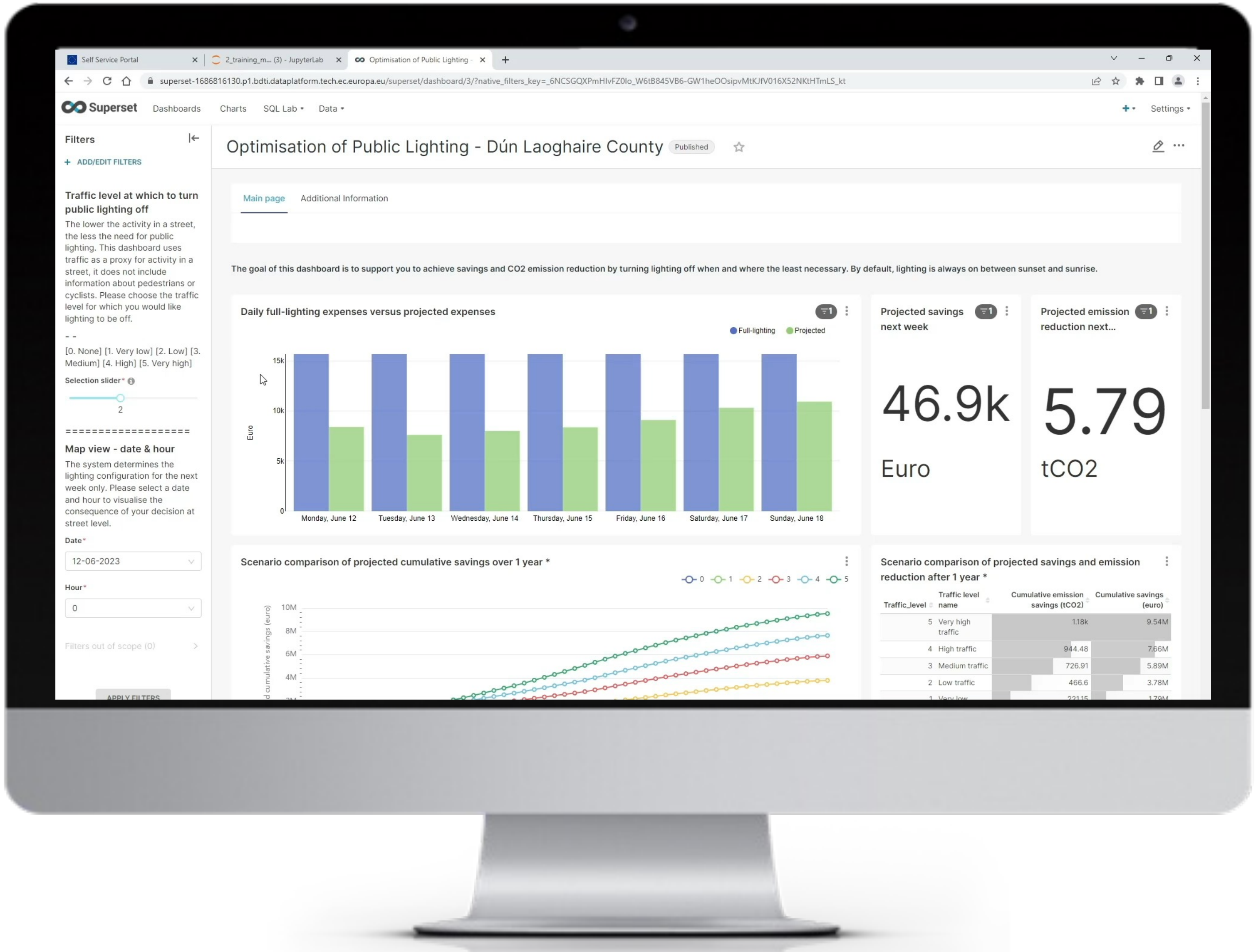
Category Lvl 0	Category Lvl 1	Expenditure (euro)
Housing and Building	RAS & Leasing Programme	25M
Environmental Services	Operation of Fire Service	19.9M
Recreation and Amenity	Outdoor Leisure Areas Operations	17.8M
Recreation and Amenity	Operation of Library and Archival Service	12.9M
Water Services	Water Supply	11.2M
Housing and Building	Maintenance & Improvement of LA Housing Units	10.9M
Housing and Building	Support to Housing Capital Prog.	8.37M
Miscellaneous Services	Administration of Rates	8.19M
Road Transport & Safety	Local Road - Maintenance and Improvement	8.04M
Development Management	Development Management	6.9M
Road Transport & Safety	Regional Road - Maintenance and Improvement	6.81M



3

Apache Superset™

Can display the created data through an interactive dashboard.
Thanks to the ML trained model, it is possible to create different simulations and visualise the outcome through different graphics


Self Service Portal
2_training_m... (3) - JupyterLab
Optimisation of Public Lighting -
superset-1686816130.p1.bdti.dataplatform.tech.ec.europa.eu/superset/dashboard/3/?native_filters_key=_6NCSGQXPmHlvFZ0lo_W6tB845VB6-GW1heOOsipvMtKJfV016X52NKtHTmLS_kt
Superset
Dashboards
Charts
SQL Lab
Data
Settings
Filters
+ ADD/EDIT FILTERS
Published
Main page
Additional Information
APPLY FILTERS

Optimisation of Public Lighting - Dún Laoghaire County

Traffic level at which to turn public lighting off

The lower the activity in a street, the less the need for public lighting. This dashboard uses traffic as a proxy for activity in a street, it does not include information about pedestrians or cyclists. Please choose the traffic level for which you would like lighting to be off.

- -

[0. None] [1. Very low] [2. Low] [3. Medium] [4. High] [5. Very high]

Selection slider*: 2

====================

Map view - date & hour

The system determines the lighting configuration for the next week only. Please select a date and hour to visualise the consequence of your decision at street level.

Date*: 12-06-2023

Hour*: 0

Filters out of scope (0)

The goal of this dashboard is to support you to achieve savings and CO2 emission reduction by turning lighting off when and where the least necessary. By default, lighting is always on between sunset and sunrise.

Daily full-lighting expenses versus projected expenses

Projected savings next week

46.9k Euro

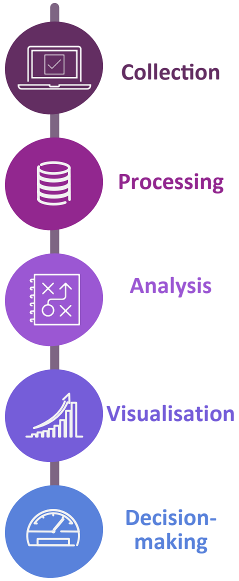
Projected emission reduction next...

5.79 tCO2

Scenario comparison of projected cumulative savings over 1 year *

Scenario comparison of projected savings and emission reduction after 1 year *

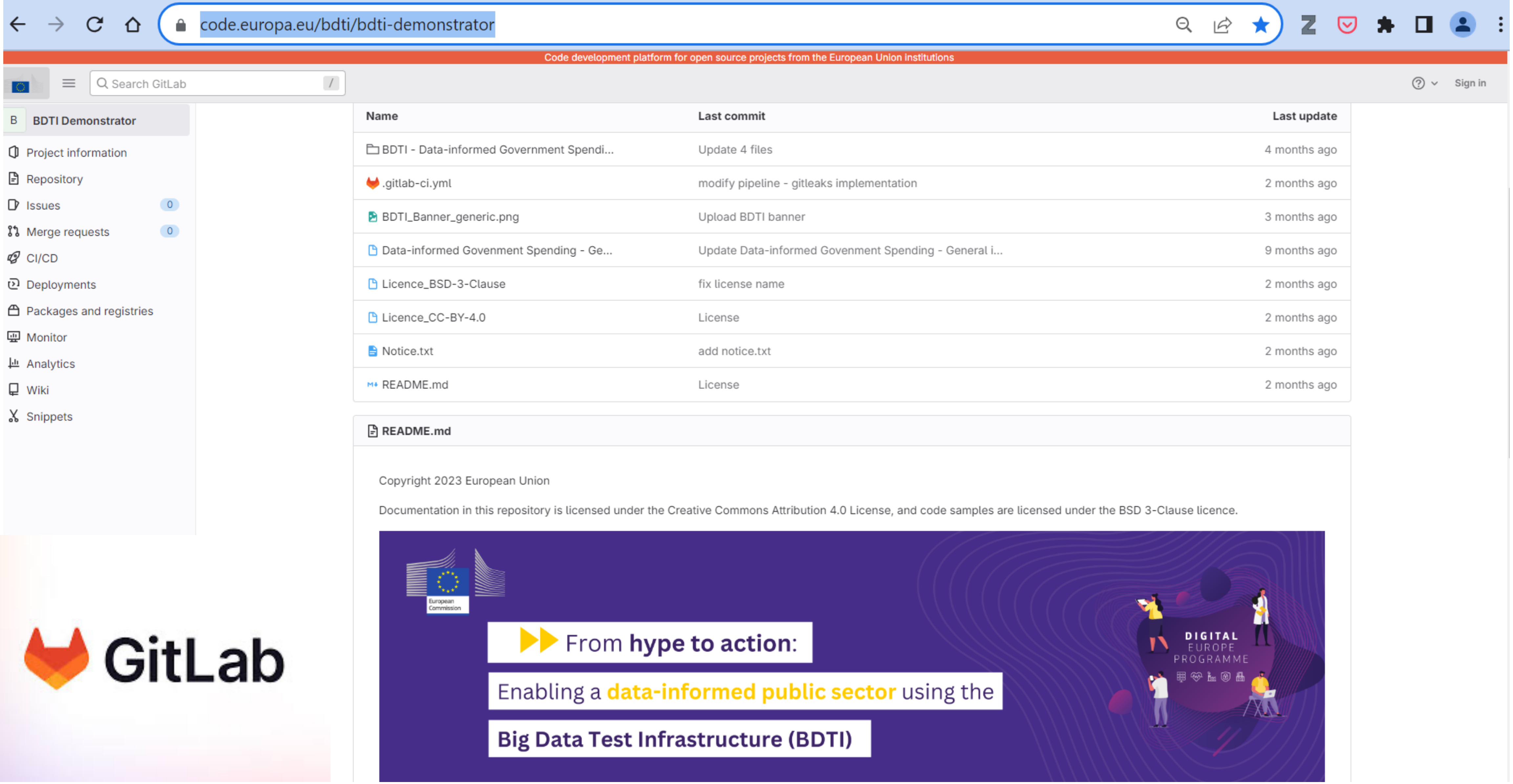
Traffic_level	Traffic level name	Cumulative emission savings (tCO2)	Cumulative savings (euro)
5	Very high traffic	1.18k	9.54M
4	High traffic	944.48	7.66M
3	Medium traffic	726.91	5.89M
2	Low traffic	466.6	3.78M
1	Very low	221.15	1.79M



Collection
Processing
Analysis
Visualisation
Decision-making

3

https://code.europa.eu/bdti/bdti-demonstrator



code.europa.eu/bdti/bdti-demonstrator

Code development platform for open source projects from the European Union institutions

Name	Last commit	Last update
BDTI - Data-informed Government Spendi...	Update 4 files	4 months ago
.gitlab-ci.yml	modify pipeline - gitleaks implementation	2 months ago
BDTI_Banner_generic.png	Upload BDTI banner	3 months ago
Data-informed Government Spending - Ge...	Update Data-informed Govenment Spending - General i...	9 months ago
Licence_BSD-3-Clause	fix license name	2 months ago
Licence_CC-BY-4.0	License	2 months ago
Notice.txt	add notice.txt	2 months ago
README.md	License	2 months ago

README.md

Copyright 2023 European Union

Documentation in this repository is licensed under the Creative Commons Attribution 4.0 License, and code samples are licensed under the BSD 3-Clause licence.

European Commission
From hype to action:
Enabling a data-informed public sector using the
Big Data Test Infrastructure (BDTI)
DIGITAL EUROPE PROGRAMME

GitLab

4

BDTI's community

- Developing the BDTI community and how can you help us

4 Who used it already?

DIGITAL EUROPE PROGRAMME

CONSELLERIA DE SANITAT (CS) - Text Mining

Conselleria de Sanitat, the Health Public Administration of the Comunidad Valenciana Regional Government, needed a tool capable of analysing and extract knowledge from the huge quantity of scientific clinical articles coming from different sources (i.e. PubMed.gov, Covid-19 related clinical articles).

GENERALITAT VALENCIANA
Conselleria de Sanitat Universal i Salut Pública

Advanced **data visualization** and **text mining** tools to help **extracting knowledge contained in the documents**, supporting clinicians and managers in their clinical practices andd day-to-day work.

EU CONVALESCENT PLASMA DATABASE – Data sharing

The European Blood Alliance is working together with the European Commission (DG SANTE) to create and manage an **EU-wide open-access platform** that collects data to support a study on **Covid-19 convalescent plasma therapy**. The aim of the study is to assess in which conditions the convalescent plasma treatment is most effective, in order to take data driven decisions on the therapy and focus the efforts of the research in the most promising directions.

eba EUROPEAN BLOOD ALLIANCE

A ready-to-use, virtual environment in which **data collected through a custom-built website** are ingested and anonymized, to be then analyzed with advanced data visualization and analytical tools. Initially, only donation data were processed, then the scope was increased to capture the **end-to-end of blood plasma, from donation to patient/clinical trial.**

CITY OF FLORENCE – Mobility data

The main goal of the Municipality is to perform a **cross correlation between the multiple datasets** available within the city to understand how people were and are moving between the different districts, to then derive precious insights about mobility the most and about **how services can be redesigned to foster cultural activities and events.**

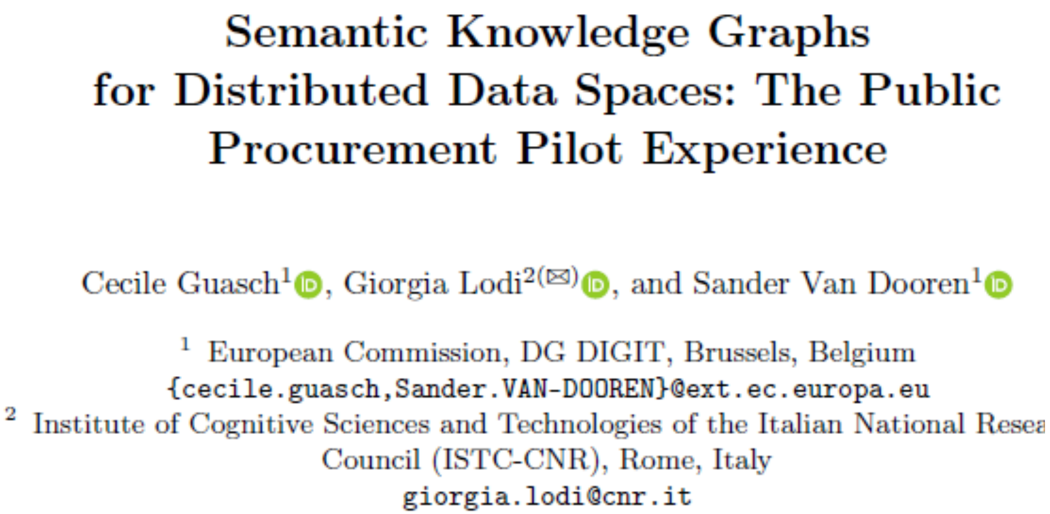
Città di FIRENZE

Predictive, descriptive and time-series analysis on multiple datasets collected **before, during and after the Covid-19 pandemic** such as: public Wi-Fi sensors, parking and geo-referenced data of people movements (i.e. tourists).

4

Who used it already?
Semantic Knowledge Graphs for Distributed Data Spaces

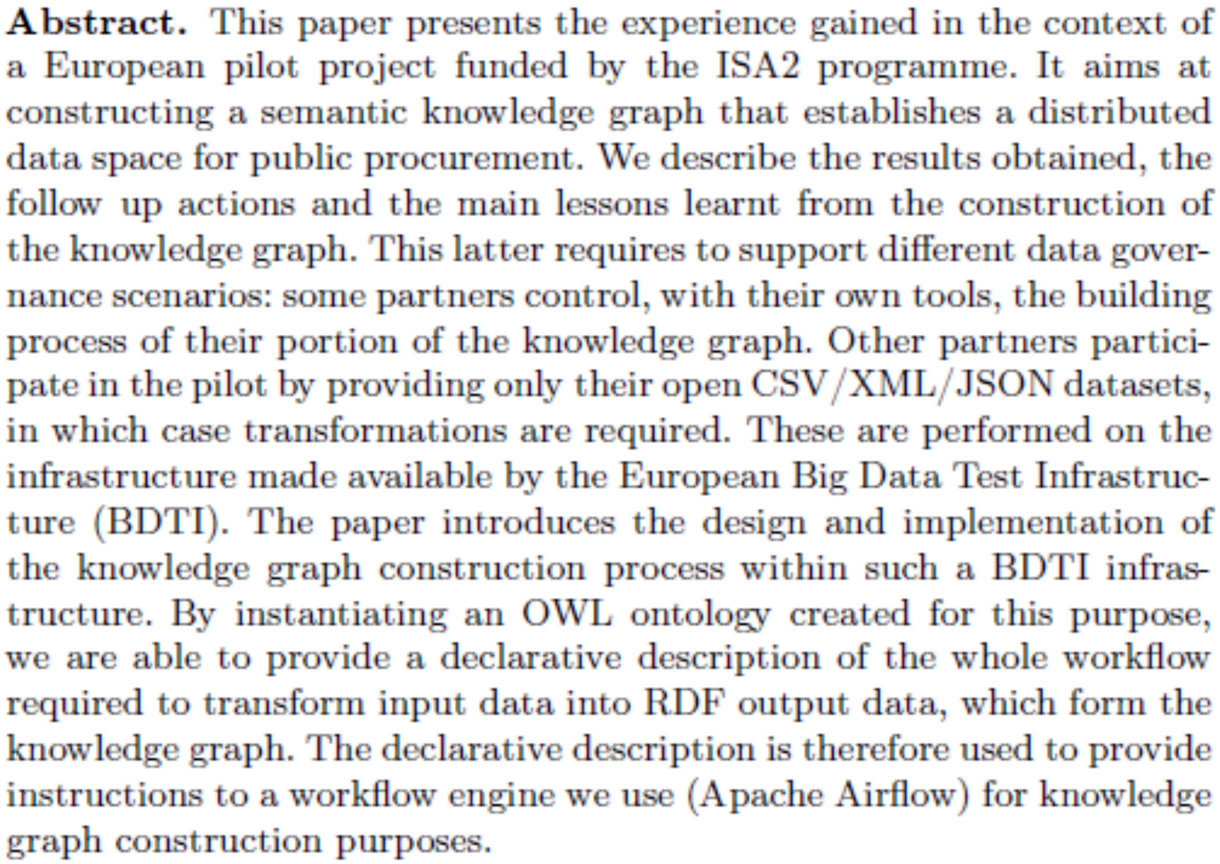
The Public Procurement Pilot Experience



Semantic Knowledge Graphs for Distributed Data Spaces: The Public Procurement Pilot Experience

Cecile Guasch[1], Giorgia Lodi[2](✉), and Sander Van Dooren[1]

[1] European Commission, DG DIGIT, Brussels, Belgium
{cecile.guasch,Sander.VAN-DOOREN}@ext.ec.europa.eu
[2] Institute of Cognitive Sciences and Technologies of the Italian National Resea
Council (ISTC-CNR), Rome, Italy
giorgia.lodi@cnr.it



Abstract. This paper presents the experience gained in the context of a European pilot project funded by the ISA2 programme. It aims at constructing a semantic knowledge graph that establishes a distributed data space for public procurement. We describe the results obtained, the follow up actions and the main lessons learnt from the construction of the knowledge graph. This latter requires to support different data governance scenarios: some partners control, with their own tools, the building process of their portion of the knowledge graph. Other partners participate in the pilot by providing only their open CSV/XML/JSON datasets, in which case transformations are required. These are performed on the infrastructure made available by the European Big Data Test Infrastructure (BDTI). The paper introduces the design and implementation of the knowledge graph construction process within such a BDTI infrastructure. By instantiating an OWL ontology created for this purpose, we are able to provide a declarative description of the whole workflow required to transform input data into RDF output data, which form the knowledge graph. The declarative description is therefore used to provide instructions to a workflow engine we use (Apache Airflow) for knowledge graph construction purposes.

Guasch, C., Lodi, G., & Dooren, S. V. (2022, October). Semantic Knowledge Graphs for Distributed Data Spaces: The Public Procurement Pilot Experience. In *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings* (pp. 753-769). Cham: Springer International Publishing. https://iswc2022.semanticweb.org/index.php/accepted-papers/

From Hype to Action

The BDTI Canva

by the BTDI Team

The BDTI Canva aims to help you build a strong data use case through a series of questions.

For more information,
visit the BDTI website

Contact us by email:
EC-BDTI-PILOTS@ec.europa.eu

Context:
Who are you? Who are your stakeholders?

Objective(s):
What is the problem you are trying to address?
What is your timeframe?

Data's added value:
Which information helps you address the problem? From which sector and or domain?

Data's availability:
Does the data you need exist?If it doesn't exist, can you collect it? From whom can you get the data you need? Can you reuse the data? What license applies to the data you'd like to use? How is the quality of the data you'd like to use? Are the different datasets interoperable? Do you know how to connect the dots?

Data's risk(s):
What could go wrong when using data to address this objective? Are there legal and ethical considerations to make? Are you dealing with personal data?

Data's processing:
What do you need to gather, process and analyze the data (i.e., tools, software, computing power, ...)? Do you already have them? If you do not, where can you get them (e.g., applying to the BDTI)?

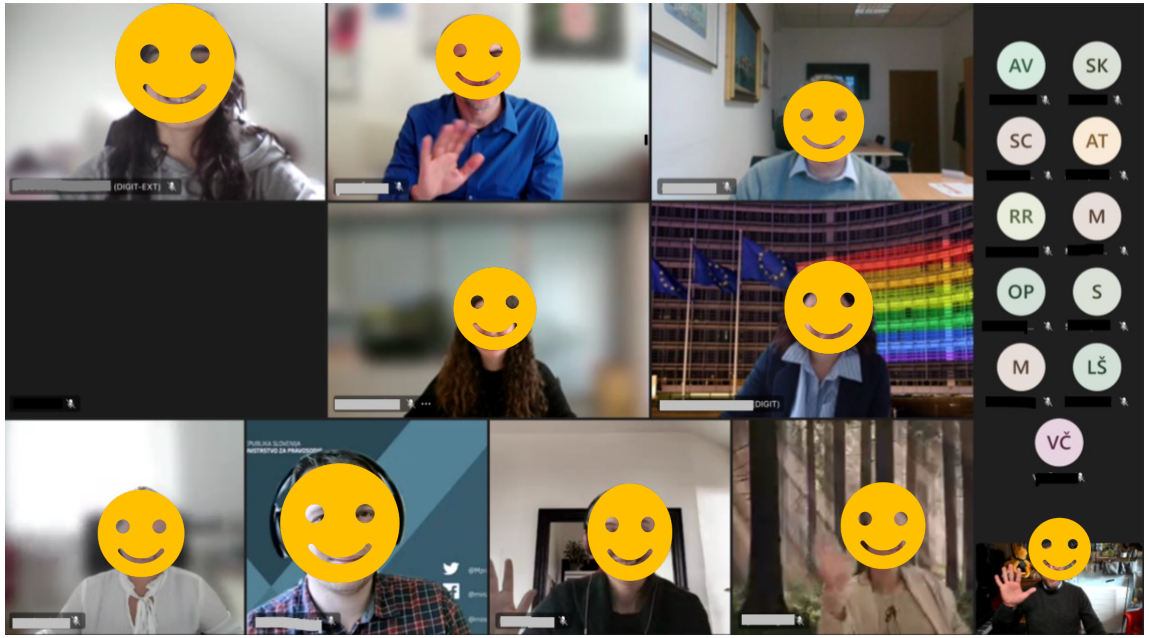
Data skills:
What data literacy and skills do you need (i.e., data engineering, data analysis, data science, data visualization)? Do you already have these available within your team/organization?

Your solution
Combine what you've learned from the elements above into a statement describing your solution

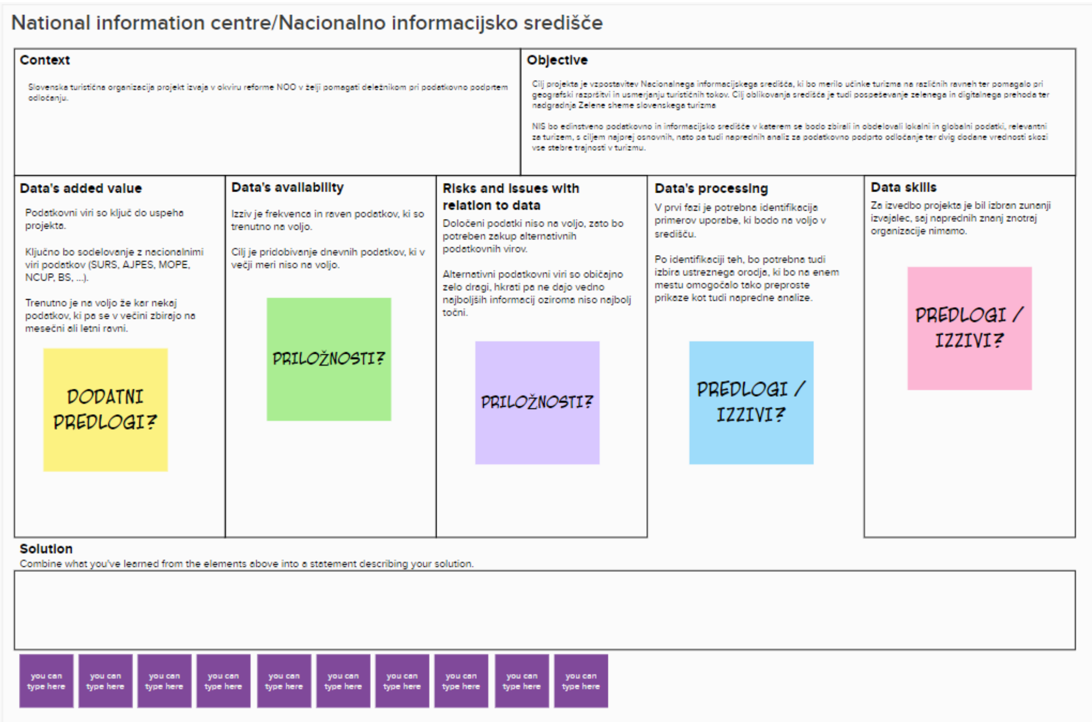
4

BDTI National Information Sessions

Goal: introduce BDTI, learn about data analytics projects, develop your data analytics community!



BDTI Information Session in Slovenia in collaboration with the **Slovenian Ministry of Digital Transformation**



National information centre/Nacionalno informacijsko središče

Context

Slovenska turistična organizacija projekt izvaja v okviru reforme NOO v želji pomagati deležnikom pri podatkovno podprtem odločanju.

Objective

Cilj projekta je vzpostavitev Nacionalnega informacijskega središča, ki bo merilo učinke turizma na različnih ravneh ter pomagalo pri geografski razpršitvi in usmerjanju turističnih tokov. Cilj oblikovanja središča je tudi pospeševanje zelenega in digitalnega prehoda ter nadgradnja Zelene sheme slovenskega turizma

NIS bo edinstveno podatkovno in informacijsko središče v katerem se bodo zbirali in obdelovali lokalni in globalni podatki, relevantni za turizem, s ciljem najprej osnovnih, nato pa tudi naprednih analiz za podatkovno podprto odločanje ter dvig dodane vrednosti skozi vse stebre trajnosti v turizmu.

Data's added value

Podatkovni viri so ključ do uspeha projekta.

Ključno bo sodelovanje z nacionalnimi viri podatkov (SURS, AJPES, MOPE, NCUP, BS, ...).

Trenutno je na voljo že kar nekaj podatkov, ki pa se v večini zbirajo na mesečni ali letni ravni.

DODATNI PREDLOGI?

Data's availability

Izziv je frekvenca in raven podatkov, ki so trenutno na voljo.

Cilj je pridobivanje dnevnih podatkov, ki v večji meri niso na voljo.

PRILOŽNOSTI?

Risks and issues with relation to data

Določeni podatki niso na voljo, zato bo potreben zakup alternativnih podatkovnih virov.

Alternativni podatkovni viri so običajno zelo dragi, hkrati pa ne dajo vedno najboljših informacij oziroma niso najbolj točni.

PRILOŽNOSTI?

Data's processing

V prvi fazi je potrebna identifikacija primerov uporabe, ki bodo na voljo v središču.

Po identifikaciji teh, bo potrebna tudi izbira ustreznega orodja, ki bo na enem mestu omogočalo tako preproste prikaze kot tudi napredne analize.

PREDLOGI / IZZIVI?

Data skills

Za izvedbo projekta je bil izbran zunanji izvajalec, saj naprednih znanj znotraj organizacije nimamo.

PREDLOGI / IZZIVI?

Solution

Combine what you've learned from the elements above into a statement describing your solution.

you can type here
you can type here
you can type here
you can type here
you can type here
you can type here
you can type here
you can type here
you can type here
you can type here

BDTI Canva used in Mural during the BDTI Information Session in Slovenia

BDTI Essentials Course – February 2024

Foundation course
6 online sessions
suitable to all
levels

Registration will be open next week

Become familiar with open-source data analytics tools

A free course helping public administrations explore BDTI delivered through a practical use case. Analysing H2020 funding allocated for research and innovation to universities across EU nations with high carbon emissions

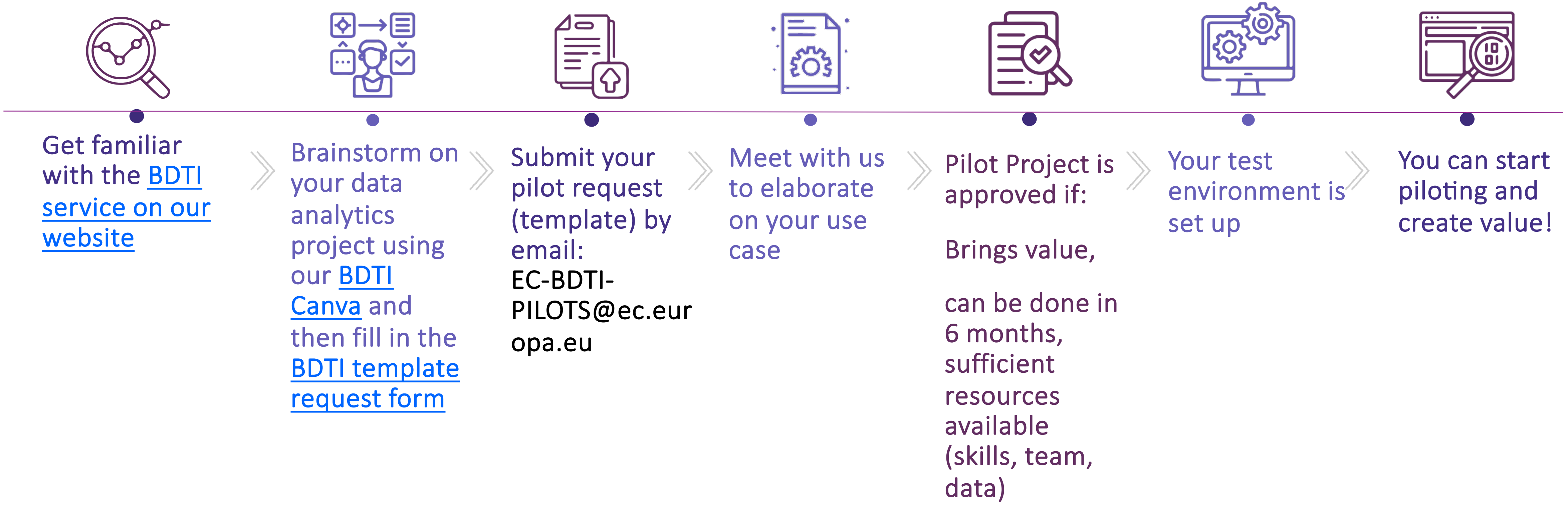
Use open-data sources for public sector innovation

Learning how to harness open data sources to address a real-world application by leveraging the resources offered by data.europa.eu

Prepare to build your own data use case

After this course, you will be ready to apply for BDTI and build a public sector data use case using the platform

How to apply:



Get familiar with the BDTI service on our website

Brainstorm on your data analytics project using our BDTI Canva and then fill in the BDTI template request form

Submit your pilot request (template) by email: EC-BDTI-PILOTS@ec.europa.eu

Meet with us to elaborate on your use case

Pilot Project is approved if:

Brings value,

can be done in 6 months, sufficient resources available (skills, team, data)

Your test environment is set up

You can start piloting and create value!

Get in touch and follow the BDTI activities

Are you working for a public administration in need of infrastructure for data analytics?

EC-BDTI-PILOTS@ec.europa.eu

Visit BDTI's website

Subscribe to BDTI's newsletter

Subscribe to BDTI's Joinup

References

Academic references:

Guasch, C., Lodi, G., & Dooren, S. V. (2022, October). Semantic Knowledge Graphs for Distributed Data Spaces: The Public Procurement Pilot Experience. In The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings (pp. 753-769). Cham: Springer International Publishing. https://iswc2022.semanticweb.org/index.php/accepted-papers/

Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, *76*(6), 928-937.

Pirog, M. A. (2014). Data will drive innovation in public policy and management research in the next decade. *Journal of Policy Analysis and Management*, 537-543.

Tan, E., & Crompvoets, J. (Eds.). (2022). *The new digital era governance: How new digital technologies are shaping public governance*. Wageningen Academic Publishers.

European Commission websites:

https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

https://digital-strategy.ec.europa.eu/en/policies/legislation-open-data

https://commission.europa.eu/publications/interoperable-europe-act-proposal_en

https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113

https://digital-strategy.ec.europa.eu/en/activities/digital-programme

https://dssc.eu/wp-content/uploads/2023/03/DSSC-Data-Spaces-Glossary-v1.0.pdf

https://digital-strategy.ec.europa.eu/en/library/staff-working-document-data-spaces