



D3.6 SYNERGY Integrated Platform & Open APIs –
Release 1.00





Digitising and transforming European industry and services: digital innovation hubs and platforms

Deliverable n°:	D3.6
Deliverable name:	SYNERGY Integrated Platform & Open APIs – Release 1.00
Version:	1.00
Release date:	02/03/2022
Dissemination level:	Public
Status:	Final
Author:	MAG, Suite5, UBITECH, KBZ, ETRA, ICCS, UCY, COBRA, VERD



Document history:

Version	Date of issue	Content and changes	Edited by
0.10	18/01/2021	Template version for DEM deliverables	Suite5
0.20	05/11/2021	Initial D3.6 Table of Contents	MAG, Suite5
0.30	01/12/2021	Contribution to section 8.2.12	COBRA
0.31	02/12/2021	Contribution to section 8.2.8	ETRA
0.32	09/12/2021	Contribution to sections 8.2.1, 8.2.2, 8.2.8, 8.2.12, 8.2.4	Suite5, ETRA, COBRA, UBI
0.33	13/12/2021	Contribution to sections 8.2.7, 8.2.3	UCY, MAG
0.34	15/12/2021	Contribution to sections 8.2.9, 8.2.5, 8.2.6, 8.2.2, 8.2.10	ICCS, VERD, Suite5
0.35	17/12/2021	Contribution to sections 8.2.11	KBZ
0.40	21/12/2021	Initial partners contributions in Sections 2-7	MAG, Suite5, UBITECH
0.50	28/01/2022	Updated partners contributions in Sections 2-7	MAG, Suite5, UBITECH
0.80	11/02/2022	Full draft available for review	MAG
0.90	28/02/2022	Final version, addressing the internal review comments	MAG, Suite5, UBITECH
1.00	02/03/2022	Final version, ready for submission	MAG

Peer reviewed by:

Partner	Reviewer
UBI	Dimitris Miltiadou
VERD	Marily Efstratiadi



Deliverable beneficiaries:

WP / Task	WP / Task	WP / Task
WP3 / T3.2-T3.5	WP5 / T5.1-T5.4	WP7 / T7.1-T7.4
WP4 / T4.1-T4.5	WP6 / T6.1-T6.4	WP8 / T8.1-T8.4



Table of contents

Executive summary.....10

1 Introduction12

 1.1 Purpose of the document 12

 1.2 Scope of the document 13

 1.3 Structure of the document 14

2 SYNERGY Integrated Platform at a Glance16

3 Data Check-in User Journey.....21

 3.1 View all Data Check-in Jobs 21

 3.2 Create a new Data Check-in Job 22

 3.2.1 Data Ingestion Configuration 24

 3.2.2 Pre-processing Rules Definition 33

 3.2.3 Define New Data Asset Profile 45

 3.3 Execute a Data Check-in Job 50

 3.3.1 Pre-processing Rules Execution 50

 3.4 Upload Data through the On-Premise Environments 53

 3.5 Manage Data Check-in Jobs 54

 3.5.1 Edit a Data Check-in Job 54

 3.5.2 Append data to a Batch File Check-in Job 55

 3.5.3 Delete a Data Check-in Job 56

 3.5.4 View the Execution Logs of a Data Check-in Job 56

 3.6 Manage Data Asset Profiles 56

 3.6.1 View all Data Asset Profiles 56

 3.6.2 Delete a Data Asset Profile 57



4	Data Search and Acquisition User Journey	58
4.1	Navigate to the SYNERGY Marketplace	59
4.2	Acquire a single Data Asset	61
4.2.1	Request a single Data Asset (Data Asset Consumer Perspective)	62
4.2.2	Review a Data Asset Request (Data Asset Provider Perspective)	63
4.2.3	Prepare a draft contract (Data Asset Provider Perspective)	64
4.2.4	Review a draft contract (Data Asset Consumer Perspective)	64
4.2.5	Negotiate a draft contract (Data Asset Consumer Perspective)	66
4.2.6	Review a revised contract (Data Asset Provider Perspective)	66
4.2.7	Settle a finalized contract (Data Asset Provider Perspective)	67
4.3	Acquire multiple Data Assets through the Cart	68
4.3.1	Request multiple Data Assets (Data Asset Consumer Perspective)	68
4.3.2	Managing the Contracts in a Bundle (Provider & Consumer Perspectives)	70
4.3.3	Settle the Contracts in a Bundle through Crypto-Currency (Consumer Perspectives)	71
4.4	Retrieve a Data Asset	72
4.4.1	View all Saved Retrieval Queries	72
4.4.2	Create a Retrieval Query	73
5	Data Analytics User Journey	76
5.1	View all Analytics Pipelines	76



5.2	Create an Analytics Pipeline	77
5.3	Configure an Analytics Pipeline	78
5.3.1	Graph View	78
5.3.2	Table View	81
5.3.3	Results View	82
5.3.4	Schedule Execution	83
5.3.5	Execution History	84
5.4	Register a pre-trained Analytics Model	84
5.5	Make a pre-trained Analytics Model or the Results of an Analytics Pipeline, available in the SYNERGY Marketplace	86
5.6	Visualize the Results of an Analytics Pipeline	90
6	Additional Platform Functionalities	91
6.1	CIM Manager	91
6.2	Manage an Organization Profile	94
6.3	Manage the User Profile	95
6.3.1	Configure the Notifications settings	95
6.3.2	Generate Access Tokens	96
6.3.3	Register an On-Premise Execution Environment	97
6.4	Manage the Organization’s Wallet	98
6.4.1	Import an Existing Wallet	98
6.4.2	Create a New Wallet	99
6.4.3	View Wallet Details	101
7	Platform Integration and Support Activities	102
7.1	Integration Approach	102
7.2	Platform Availability	104
7.3	Platform Support Mechanisms	104



8	SYNERGY Baseline Analytics	106
8.1	Baseline Data Analytics	106
8.2	Baseline Data Analytics Updates	108
8.2.1	AC consumption for different comfort levels	108
8.2.2	AC consumption flexibility forecasting	112
8.2.3	Anomaly Detection in household energy consumption	115
8.2.4	Building Energy Demand Forecasting	120
8.2.5	Prediction of day-ahead demand flexibility at building level	125
8.2.6	Prediction of generation flexibility at DER level - short-term	130
8.2.7	Anomaly detection in energy demand at building level	135
8.2.8	Outlier detection in energy demand	140
8.2.9	Prediction of peak and average energy generation at portfolio level in specific TSO/DSO areas (over a year)	146
8.2.10	Prediction of storage flexibility	149
8.2.11	Malfunction Duration Prediction in PV energy generation	152
8.2.12	Clustering of PV malfunctions/inefficiencies	156
9	Conclusions & Next Steps	162
	Annex I: References	164
	Annex II: Stress Activities Outline.....	165



Abbreviations and Acronyms

Acronym	Description
AI	Artificial Intelligence
API	Application Programming Interface
CIM	Common Information Model
CSV	Comma-Separated Values
DL	Deep Learning
DLT	Distributed Ledger Technologies
DoA	Description of Action (annex I of the Grant Agreement)
DCJ	Data Check-in Job
JSON	JavaScript Object Notation
ML	Machine Learning
OPE	On-Premise Environment
SASL	Simple Authentication and Security Layer
TSV	Tab-Separated Values
XML	Extensible Markup Language
WP	Work Package



Executive summary

This document D3.6 “SYNERGY Integrated Platform & Open APIs – Release 1.00” presents the first, official release of the SYNERGY Integrated Platform that integrates the release 1.00 of the services bundles developed under WP3 “End-to-end Interoperable Big Data Management Platform”, and WP4 “Big Data Analytics and Data Sharing Mechanisms”, in order to deliver a seamless user experience. These services bundles include the Data Collection Services Bundle, the Data Security Services Bundle, the Data Sharing Services Bundle, the Data Matchmaking Services Bundle, the Data Analytics Services Bundle, the Data Governance Services Bundle, the Data Storage Services Bundle, and the Platform Management Services Bundle.

The purpose of this deliverable is to document the functionalities of Release 1.00 of the SYNERGY Integrated Platform that brings added value to the different electricity data value chain stakeholders for efficient data management, trusted data sharing and insightful data analytics. To this end and building on the beta release that was documented in the SYNERGY Deliverable D3.4, the set of core functionalities that are offered by the SYNERGY Integrated Platform are described from a user-oriented perspective across three core user journeys: (a) the data check-in user journey that describes the process that data asset providers need to follow to make their assets available in the SYNERGY platform; (b) the data asset search and acquisition user journey which describes the process of searching, exploring, and acquiring data assets (that belong to other organizations) through the SYNERGY Data & AI Marketplace; and (c) the data analytics user journey that describes the workflow of designing data analytics pipelines and their execution, in order for the data asset providers / consumers to gain valuable insights regarding their own and/or acquired data assets. Moreover, the SYNERGY Platform enables organization-based access, prepares the ground for single sign-on functionalities with the SYNERGY energy apps and allows the lifecycle management of the SYNERGY Common Information Model (CIM).

The current deliverable also presents twelve (12) additional baseline analytics solutions that have been integrated in the SYNERGY Integrated Platform and complement the portfolio of pre-trained analytics for Demand Forecasting, Generation Forecasting, Occupants’ Behaviour and Comfort Profiling, Flexibility Forecasting, and Predictive and Preventive Maintenance, that was initially presented in the SYNERGY Deliverable D4.2.

As the platform development activities remain ongoing, additional refinements and enhancements will be analysed and introduced as needed in release 2.00 (expected on M36),



according to the SYNERGY integration plan and requirements, as well as to the feedback that is continuously collected by the SYNERGY energy apps (in WP5-WP7), the demonstration activities (in WP8), and the living lab activities (engaging external stakeholders) in WP9.



1 Introduction

1.1 Purpose of the document

The SYNERGY Deliverable D3.6 “SYNERGY Integrated Platform & Open APIs – Release 1.00” accompanies the first official release of the SYNERGY Integrated Platform and intends to document its functionalities from a user-oriented perspective. The SYNERGY Integrated Platform is mainly delivered under task T3.5 “Platform and Services Bundles Continuous Integration and SYNERGY Open APIs Delivery”, in collaboration with tasks T3.2 “Platform Backbone Infrastructure, On-Premise and Secure Experimentation Playground Data Containers Development”, T3.3 “Core Big Data Ingestion, Curation and Management Services”, T3.4 “Data Assets Security, Encryption and Privacy Mechanisms”, T4.1 “Big Data Analytics Workbench and Jobs Execution Engines”, T4.4 “Blockchain-enabled, Trusted Multi-Party Data Sharing Services”, and T4.5 “Big Data Exploration and Matchmaking Services”. The documentation presented in D3.6 focuses on the end-to-end, integrated functionalities delivered through the different data services bundles, namely the Data Collection Services Bundle, the Data Security Services Bundle, the Data Sharing Services Bundle, the Data Matchmaking Services Bundle, the Data Analytics Services Bundle, the Data Governance Services Bundle, the Data Storage Services Bundle, and the Platform Management Services Bundle. Specifically, this deliverable focuses on providing a thorough description of the workflows and their corresponding functionalities supported by the current release of the SYNERGY Platform, and any other additional platform functionality that was developed under WP3 “End-to-end Interoperable Big Data Management Platform” and WP4 “Big Data Analytics and Data Sharing Mechanisms” as reported in Release 1.00 of all services bundles that was documented in the SYNERGY Deliverables D3.5 “Data Collection, Security, Storage, Governance & Management Services Bundles - Release 1.00” and D4.3 “SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles - Release 1.00”. Essentially, this documentation provides a detailed step-by-step explanation of the various developed platform functionalities and associated workflows supported in Release 1.00 of the SYNERGY Integrated Platform & Open APIs.

In addition, in collaboration with Task T4.2 “Data Analytics Algorithms Baseline Definition”, this deliverable reports on the progress of the pre-trained analytics solutions that are integrated in the SYNERGY Platform in order to address specific problems of the electricity data value chain stakeholders.



1.2 Scope of the document

The scope of this deliverable is to document the Release 1.00 of the SYNERGY Integrated Platform, building on the initial, planned functionality of the beta version (documented in the SYNERGY Deliverable D3.4), and introducing the on-demand APIs that provide access to selected data assets that can be legitimately uploaded or retrieved by external legacy systems of the energy stakeholders or other third parties, according to their needs. More precisely, the scope of this document is:

- To present the first official version of the SYNERGY Integrated Platform including a thorough description of the platform’s core functionalities. These functionalities are described by three different core user journeys namely: (a) the data check-in user journey that denotes the process of uploading data assets to the platform; (b) the data search and acquisition user journey that involves any process related to data asset search, exploration and acquisition; and (c) the data analytics user journey that describes the process of creating and executing data analytics pipelines and visualizations to gain valuable insights from certain data assets.
- To outline the additional functionalities that are provided along with the core platform’s functionalities, to support different tasks related to the common information model lifecycle management, user and organization profile management, and wallet creation.
- To define how the open APIs that are customized per stakeholder and data asset are created and delivered along with the SYNERGY Integrated Platform. The process of generating customized, open APIs is available to any SYNERGY stakeholder to support the energy-related applications that are to be developed.
- To document the baseline analytics solutions for Demand Forecasting, Generation Forecasting, Occupants’ Behaviour and Comfort Profiling, Flexibility Forecasting, and Predictive and Preventive Maintenance, that complement the draft release (documented in the SYNERGY Deliverable D4.2).
- To perform and summarize the initial stress testing activities for the SYNERGY Platform.

The first official version of the SYNERGY Integrated Platform provides a well-documented description of the platform functionalities designed and developed under WP3 and WP4, and described in D3.5 “Data Collection, Security, Storage, Governance & Management Services Bundles – Release 1.00”, D3.4 “SYNERGY Integrated Platform – Beta Release”, and D4.3



“SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles – Release 1.00”. Moreover, the development of the SYNERGY Integrated Platform is properly aligned to the technical requirements and use cases of the SYNERGY project as documented in D2.2 “End-user and Business requirements analysis for big data-driven innovative energy services and ecosystems”, as well as the SYNERGY Architecture defined in D2.7 “SYNERGY Framework Architecture including functional, technical and communication specifications v2”.

1.3 Structure of the document

The structure of the remainder of this document is organized as follows:

- Section 2 provides a high-level view of the SYNERGY Integrated Platform, describing the main components and services bundles that are already integrated within the platform.
- Section 3 presents the data check-in user journey that guides the energy data value chain stakeholders, acting as data asset providers, through the data collection and management functionalities including both the design and the execution phases.
- Section 4 presents the data asset search and acquisition user journey that guides the energy data value chain stakeholders, acting as data consumers, through the search and acquisition functionalities that are available in the SYNERGY Data & AI Marketplace.
- Section 5 presents the data analytics user journey that the SYNERGY Integrated Platform offers to the energy data value chain stakeholders, acting as data providers or consumers, for designing and executing data analytics pipelines in order to gain valuable insights and visualizations for a particular data asset.
- Section 6 presents all the additional functionalities that are provided along with the core functionalities of the SYNERGY Integrated Platform. These functionalities support various management mechanisms for platform’s users, organization profiles, and the lifecycle of the common information model.
- Section 7 summarises the overall integration and support approach followed in the SYNERGY Platform.
- Section 8 documents the additional pre-trained analytics solutions that have been integrated in the SYNERGY Platform.



- Section 9 concludes this deliverable and provides future steps that are to be followed in the project.
- Annex I includes the list of relevant references.
- Annex II documents the stress testing activities that have been performed so far in the SYNERGY Platform.



2 SYNERGY Integrated Platform at a Glance

In accordance with the integration plan that was presented in D3.4, the SYNERGY Integrated Platform on its first official release (Release 1.00) has delivered the planned functionalities for the Cloud Platform (including the Secure Experimentation Playgrounds that are spawn per organisation) and the Server and Edge On-Premise Environments. The core workflows concerning data check-in, data search and sharing, and data analytics, are supported as described in the documentation of Release 1.00 of the various Data Services Bundles (as described in D3.5 and D4.3). An overview of the core functionalities associated with each services bundle is depicted in Figure 1. The features delivered in this release appear as underlined in the figure.

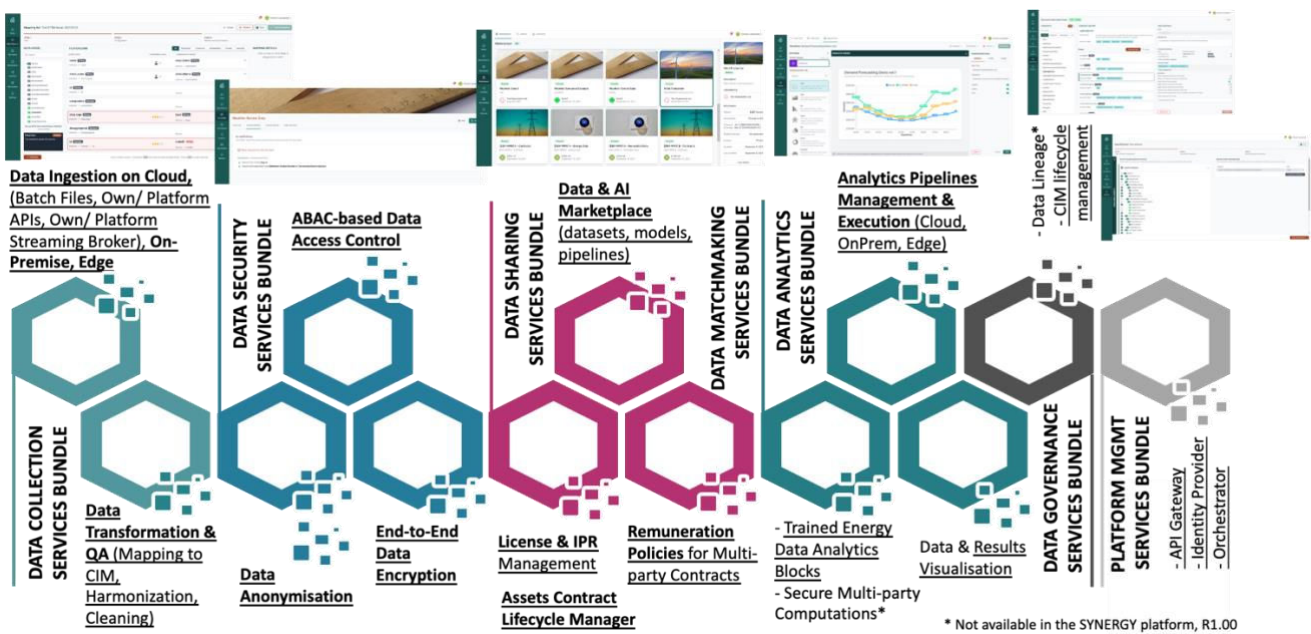


Figure 1: SYNERGY Services Bundles

In more detail, the main functionalities delivered through the SYNERGY Platform release 1.00 towards the different electricity data value chain stakeholders who at any moment may assume the role of data asset providers and / or data asset consumers, are briefly described in Table 1 below.

Table 1: SYNERGY Beta Platform - Delivered Functionalities

Data Collection Services Bundle
✓ Configuration of the data check-in process by the data provider, and its proper execution in the SYNERGY Cloud Infrastructure and / or the On-Premise Environment (in its server and edge editions).
✓ Configuration and execution of different data ingestion, mapping and transformation and cleaning services that are invoked as-needed to appropriately handle batch, near real-time (through push and pull mechanisms) and streaming data collection.
✓ Appending data in the case of batch file data upload in data check-in jobs.
✓ Revision of a failed step in a data check-in job that has not been successfully executed.
✓ Detailed execution logs for each data check-in job.
✓ Improvements in the overall performance of the data collection services.
Data Security Services Bundle
✓ Configuration of data anonymization for privacy and confidentiality preservation purposes (from an individual or business perspective).
✓ End-to-end data encryption for securing transfer from the Server On-Premise Environment to the Cloud Platform.
✓ Enforcement of access policies per data asset (datasets, models, results) to control who can or cannot view a data asset in the SYNERGY Marketplace.
Data Matchmaking Services Bundle
✓ Navigation to the Data & AI Marketplace (including datasets, models, results), search and exploration over assets that data consumers are eligible to view and potentially acquire (based on the applicable access policies).
✓ DLT-based acquisition contracts lifecycle management (including preparation, negotiation, and activation) for legitimately exchanging data assets (datasets, models, results) in an immutable manner.
✓ DLT-based derivation contracts lifecycle management (including preparation, negotiation, and activation) for allowing the original data asset providers to approve sharing a derivative data asset in the Marketplace.
✓ Simultaneous acquisition of multiple data assets through the cart functionality.
✓ Crypto-currency payments for acquisition contracts, including remuneration of all involved parties (based on the applicable derivation contracts).
Data Analytics Services Bundle
✓ Design of data analytics pipelines, along with different data manipulation configuration, the basic and baseline (pre-trained) machine learning and deep learning algorithms configuration and the visualization/results configuration.
✓ Execution of the designed and configured data analytics pipelines in both the Secure Experimentation Playgrounds and the Server/Edge On-Premise Environments.
✓ Registration of trained models created within or outside the SYNERGY Platform.
✓ Improvements in the overall performance of the data analytics services. Support for additional open-source libraries/frameworks.
Data Storage Services Bundle
✓ Storage of: (a) data assets and their accompanying metadata, along with their indexing; (b) data check-in/analysis jobs-related configuration and retrieval-related configuration; (c) algorithms, models and pipelines; (d) contracts' ledger in DLT; (e) log-related information for the SYNERGY platform operation and usage.
✓ Secure, encrypted storage of sensitive data

Data Governance Services Bundle
✓ Coordination and end-to-end management of the data across all layers of the SYNERGY platform.
✓ Lifecycle management of the Common Information Model (presented in D3.1)
Platform Management Services Bundle
✓ Retrieval of user-defined data (datasets, analytics pipelines results) through custom API endpoints exposed on-demand by the SYNERGY Platform
✓ Management of the underlying platform’s resources.
✓ Security and authentication aspects of the platform under the SYNERGY realm. Support for Single Sign-On of the SYNERGY energy apps.
✓ Management of platform and email notifications. Availability of digest daily/weekly notifications.

The status of integration among the different SYNERGY platform components is depicted in the following table. Such interactions essentially remain the same for the Cloud Platform (including the Secure Experimentation Playgrounds that are spawn per organisation) and the On-Premise Environments (in their server and edge edition). Obviously, the integration of the Data Storage Services Bundle has been ensured and it effectively communicates with all components and services in the SYNERGY Platform.

Table 2: SYNERGY Components Integration Status

Component	Related Components with which there is interaction/integration	Status in SYNERGY Platform Release 1.00
Data Collection Services Bundle		
Data Handling Manager	Matching Prediction Engine	Integration Completed
	Access Policy Engine	Integration Completed
	Master Controller	Integration Completed
Matching Prediction Engine	Mapping & Transformation Service	Integration Completed
	CIM Manager	Integration Completed
Data Ingestion Service	Master Controller	Integration Completed
Mapping & Transformation Service	Master Controller	Integration Completed
Cleaning Service	Master Controller	Integration Completed
Data Security Services Bundle		
Anonymisation Service	Master Controller	Integration Completed
Encryption Engine	Master Controller	Integration Completed
	Wallet Manager	Partial Integration
Access Policy Engine	Data Handling Manager	Integration Completed
	Data & AI Marketplace	Integration Completed
	Analytics Workbench	Integration Completed
	API Gateway	Integration Completed
Data Sharing Services Bundle		
Data & AI Marketplace	Query Builder	Integration Completed
	Matchmaking Engine	Integration Completed
	Contract Lifecycle Manager	Integration Completed



Contract Lifecycle Manager	Remuneration Engine	Integration Completed
	Notifications Engine	Integration Completed
	Master Controller	Integration Completed
Remuneration Engine	Wallet Manager	Integration Completed
Wallet Manager	Encryption Engine	Partial Integration
	Remuneration Engine	Integration Completed
Data Matchmaking Services Bundle		
Query Builder	Data & AI Marketplace	Integration Completed
	Matchmaking Engine	Integration Completed
	API Gateway	Integration Completed
Matchmaking Engine	Query Builder	Integration Postponed for the upcoming releases
	Data & AI Marketplace	Integration Postponed for the upcoming releases
Data Analytics Services Bundle		
Analytics Workbench	Visualization & Reporting Engine	Integration Completed
	Data & AI Marketplace	Integration Completed
	API Gateway	Integration Completed
Visualization & Reporting Engine	Service Secure Results Export Service	Integration Completed
Data Manipulation Service	Master Controller	Integration Completed
Analytics Execution	Master Controller	Integration Completed
Service Secure Results Export Service	Master Controller	Integration Completed
Data Governance Services Bundle		
Master Controller	Access Policy Engine, Contract Lifecycle Manager, Data Ingestion Service, Mapping & Transformation Service, Cleaning Service, Anonymisation Service, Encryption Engine, Data Manipulation Service, Analytics Execution, Service Secure Results Export Service, Data Handling Manager, Analytics Workbench	Integration Completed for all components
Data Lineage Service	Master Controller	Planned for Release 2.00
CIM Manager	Matching Prediction Engine	Integration Completed
	Query Builder	Integration Completed
Platform Management Services Bundle		
Resources Orchestrator	Master Controller	Integration Completed
Notifications Engine	Master Controller	Integration Completed
	Contract Lifecycle Manager	Integration Completed
Security, Authentication & Authorisation Engine	All Components & Services	Integration Completed
Platform Analytics Engine	Data Handling Manager, Data & AI Marketplace, Contract Lifecycle Manager	Planned for Release 2.00
API Gateway	Access Policy Engine	Integration Completed
	Contract Lifecycle Manager	Integration Completed
	Query Builder	Integration Completed
	Service Secure Results Export Service	Integration Completed





3 Data Check-in User Journey

This section describes the data check-in user journey including the various steps a data asset provider needs to follow for uploading data on the SYNERGY Platform. In the data check-in workflow (depicted in Figure 2), the data asset provider starts by creating a new data check-in job, configures the data ingestion process that is responsible to load the data on the SYNERGY Platform, defines the pre-processing rules that are to be applied on the data asset, defines the data asset’s metadata and licensing details, and finally executes the data check-in job that triggers the data loading to the SYNERGY Platform.

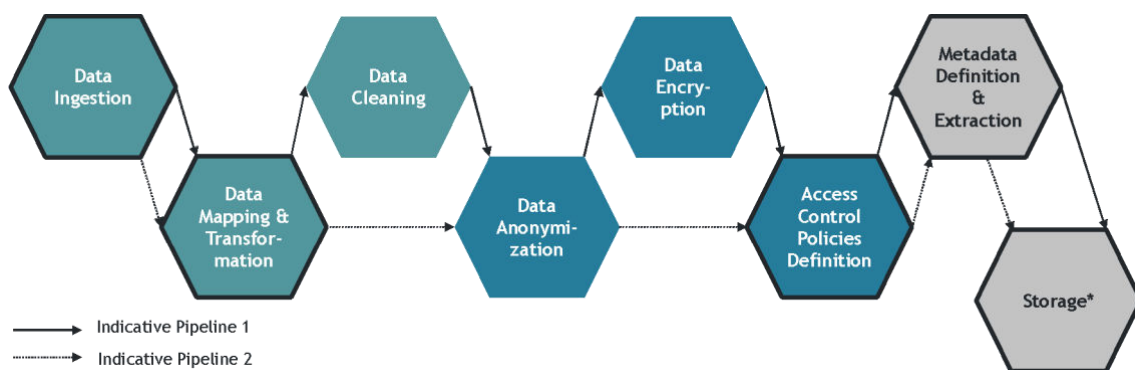


Figure 2: Data Check-in Pipelines

It needs to be noted that there is a distinct separation between the creation and configuration (design) phase of the data check-in job as described in Section 3.1, and the actual execution of the configured data check-in job by the SYNERGY Platform as described in Sections 3.3 and 3.4.

3.1 View all Data Check-in Jobs

A summary list including all the created/configured data check-in jobs appears in the Data Check-in Jobs view as shown in Figure 3. The data asset provider may view all the details regarding the data check-in jobs in a list view, including the job status and available actions depending on the job status. The date of creation and the user (within the organization) that has created the data check-in job, as well as the status of execution for each data check-in job, are also presented. In particular, when the data check-in job is already executed, a tick mark, cross mark, or exclamation mark will appear next to the corresponding data check-in step (i.e. Harvester, Mapping, Cleaning, Anonymiser, Encryption, Loader) if the execution for this step was successful, failed, or not processed in case of wrong configuration, respectively, for the steps

that are relevant for the specific job. In the case where the data check-in job is not executed yet, but instead it is in the queue for execution, then a clock icon will appear denoting an execution pending status.

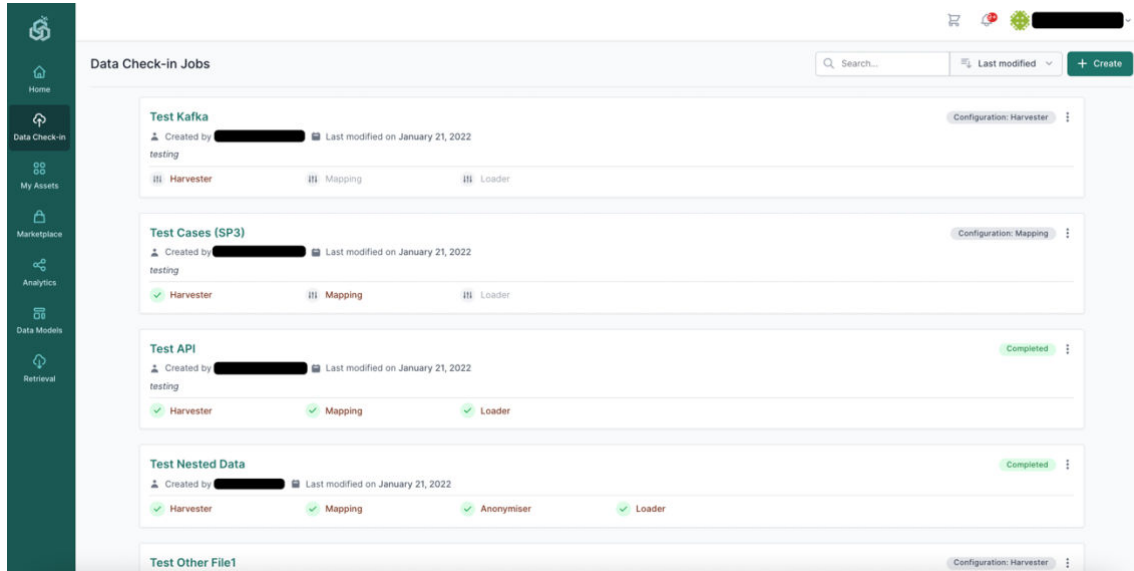


Figure 3: View All Data Check-in Jobs

3.2 Create a new Data Check-in Job

The first step of the data check-in workflow is the creation and configuration of a new data check-in job. The data asset provider may create a new data check-in job by selecting the Create button at the Data Check-in Jobs page which lists all the created data check-in jobs, as depicted in Figure 3. First, the data asset provider needs to provide basic information (i.e., title, and description) about the data check-in job in the Data Check-in Job Details section, as shown in Figure 4. Additionally, the data asset provider needs to select the pre-processing rules, that are to be applied on the loaded data, including Mapping, Cleaning, Anonymisation and Encryption. It needs to be noted that the selection of the Mapping step activates the selection of the other pre-processing steps (i.e., Cleaning, Anonymisation, Encryption). The final step for the creation of the data check-in job, is to select the environment where the data check-in job is to be executed. In particular, the data asset provider may select either Cloud Execution to run the data check-in job on the cloud, or On-premise Execution to run the data check-in job on-premise through a registered On-Premise Environment (installed in a server or in the edge, i.e. a gateway). Next, the data asset provider finalizes the creation of a data check-in job by selecting

the Create button at the top right part of the page. Finally, the data check-in job is created and shown in the list of Data Check-in Jobs page (as depicted in Figure 3).

The data asset provider can proceed to configure the different pre-processing rules that were selected before. First, it is mandatory to configure the Harvester step that is pre-selected by default. The configuration can be done by selecting the appropriate data check-in job from the list, and then by selecting the Harvester option. The Harvester configuration determines the data ingestion method that is to be followed, which sets the way of loading the data on the SYNERGY Platform. As depicted in Figure 5, the data provider may select the data loading method as follows:

- File Upload Method, allowing data asset providers to upload files in different formats (i.e., CSV, JSON, XML, and other formats) containing batch historical and operational data.
- Application's Own (External) API Method, which supports the use of an external API, owned by the data provider or any other third party, for importing the API request results.
- Platform's (Internal) API Method, which allows data uploading to API endpoints provided by the SYNERGY Platform and customized to the data that are to be pushed.
- Platform's (Internal) Kafka PubSub Method, which can be configured to upload streaming data to the Kafka Publish-Subscribe (PubSub) mechanism provided by the SYNERGY Platform.
- Application's Own (External) Kafka PubSub Method, allowing data providers to use their own Kafka PubSub mechanism and grant access to the SYNERGY Platform to collect the streaming data they publish in a specific topic.

By selecting the “Finalize Configuration” button, the corresponding configuration page depending on the selection will appear.



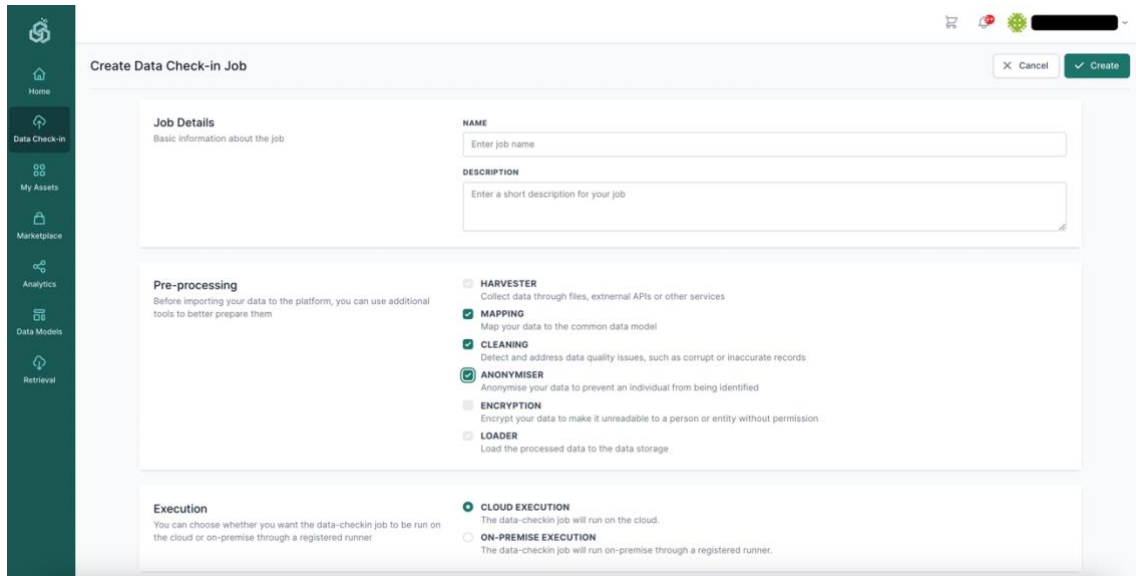


Figure 4: Create a new Data Check-in Job (DCJ)

3.2.1 Data Ingestion Configuration

3.2.1.1 File Upload Method

Once the data asset provider has selected the data upload option (shown in Figure 5), the Data Ingestion configuration page appears as depicted in Figure 6.

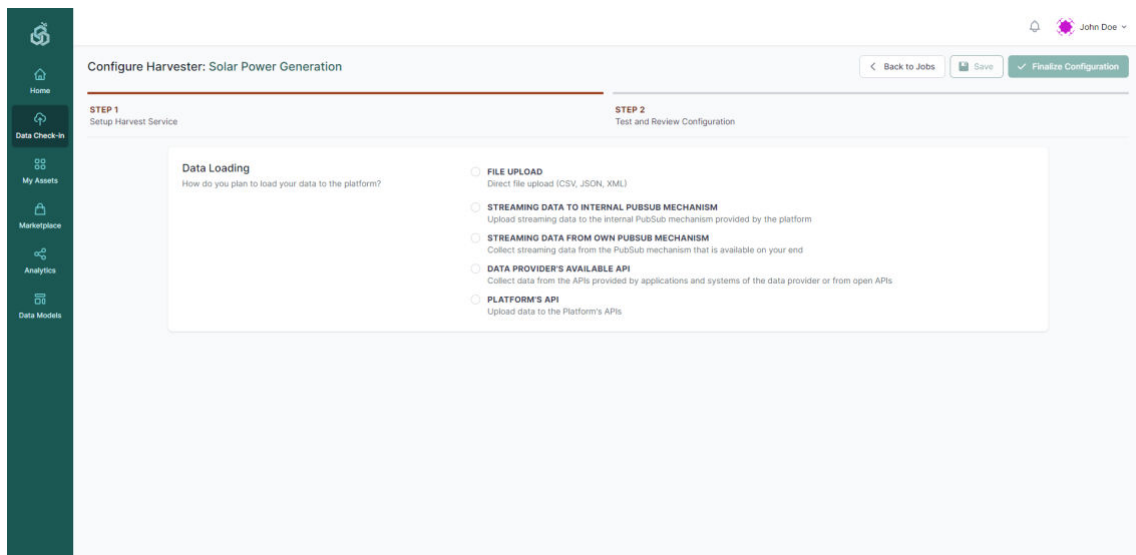


Figure 5: Data loading method

In the configuration page, the file format of the data that are to be uploaded needs to be selected. Following that, the data asset provider should upload a sample file including an indicative number of entries from the whole data asset (applies to CSV, TSV, JSON, and XML files

only), and the actual file that is to be uploaded. In the case that the data provider selects not to upload the data directly in the cloud, the location of the file where it will be stored locally and retrieved from the On-Premise Environment, needs to be provided accordingly.

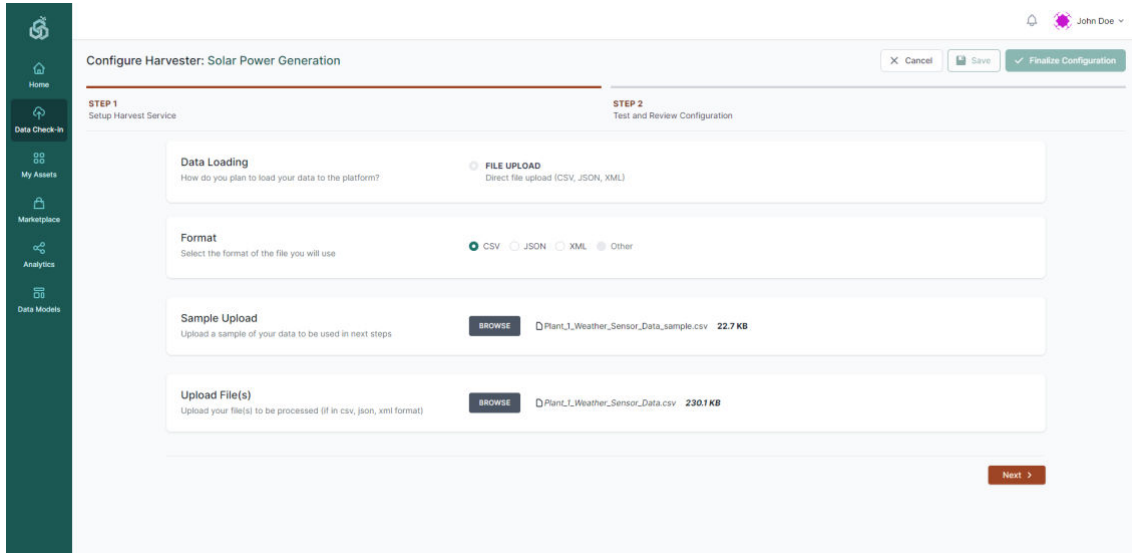


Figure 6: Data Check-in – File Upload Method - Setup Harvest Service (Step 1)

By selecting the Next button, the data asset provider may view the details of the data sample that was uploaded in a tabular view, or tree view for non-flat formats, as depicted in Figure 7. Finally, the data provider may save the configuration provided so far or finalize the configuration button that essentially completes the Harvester configuration.

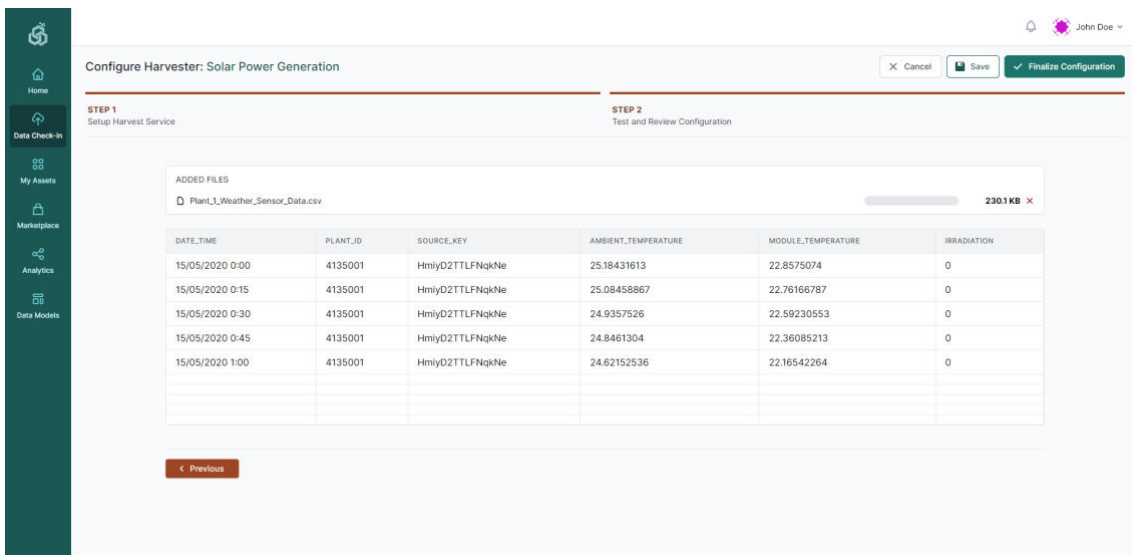


Figure 7: Data Check-in - File Upload Method - Review (Step 2)

3.2.1.2 Application’s own (external) API method

Once the data asset provider has chosen to ingest data from an own (external) API that is exposed by his/her organisation’s systems, the corresponding configuration page will appear as shown in Figure 8. Initially, the data asset provider is asked to select the format of the API response (e.g. JSON or XML), and the authentication details regarding the API access policies (e.g. None, if no authentication is needed; Bearer if authentication that involves security tokens is needed; and Custom, if the external API support the use of a custom URL for authentication). Following that, more details need to be provided, depending on the authentication type selection (e.g. for Custom Authentication, the Authentication URL and the Authentication Query Body, in order to retrieve the Access Token, allowing the data asset provider to test the authentication policies inserted, by selecting the Test Login button). Then, the full API path including the base URL, along with the appropriate method (i.e., GET, POST, PUT), need to be inserted. If a POST method is selected, the query body of the request needs to be also provided.

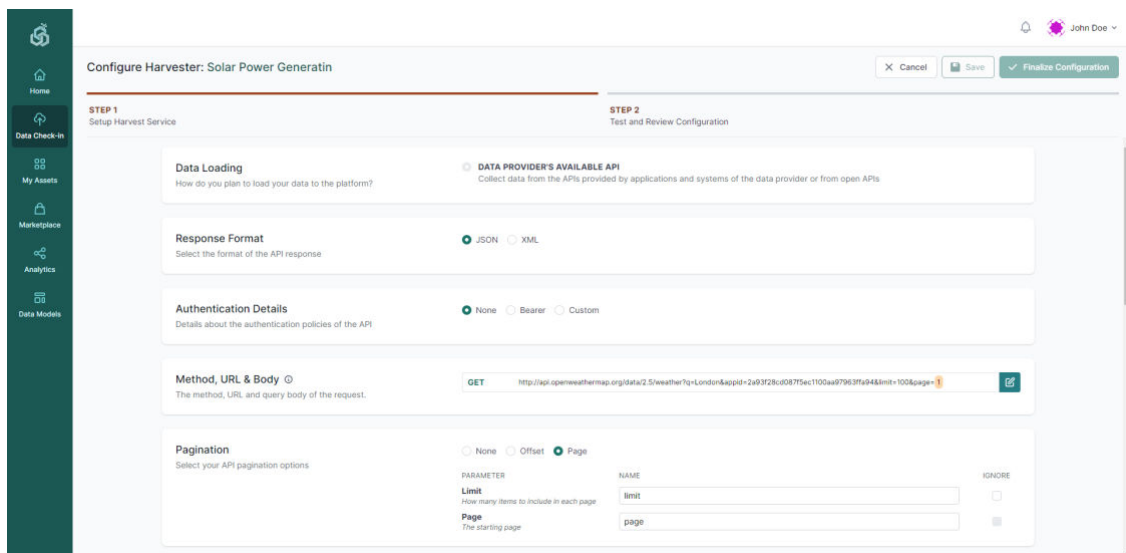


Figure 8: Data Check-in - API (external) – Setup Harvest Service (Step 1a)

Once the full API path is provided, the corresponding request parameters will appear automatically in the Request Parameters section where one can edit them, remove them, or even add new query parameters, as depicted in Figure 9. Different options for pagination of the API responses, such as offset or page, are also provided to define how paginated API responses should be handled. By inserting the API method URL and query body, the corresponding request parameters will appear in the Request Parameters sections, where the data asset provider may also insert additional request parameters. In addition to the request parameters, the data asset

provider may also add extra headers to the API calls by selecting the Add Header button. It needs to be noted that any request parameter and extra header can be treated as sensitive (that essentially means that it is stored in an encrypted form), depending on the data asset provider’s preference.

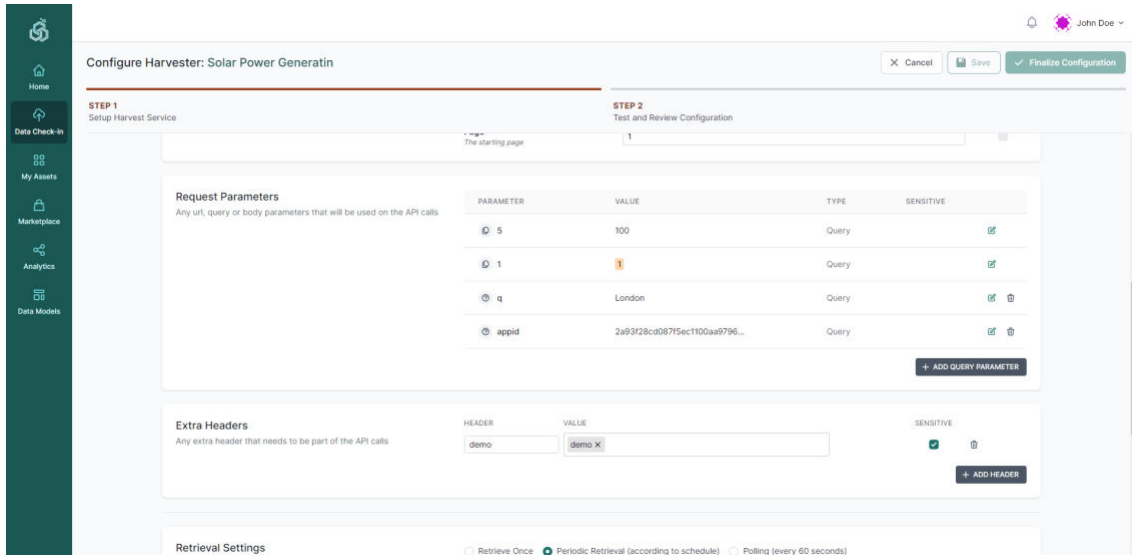


Figure 9: Data Check-in - API (external) – Setup Harvest Service (Step 1b)

Another important step of the external API data ingestion method is to define the retrieval settings regarding the schedule and periodicity of the data ingestion that the SYNERGY Platform should initiate to retrieve data from the specific API, as shown in Figure 10. In particular, the data asset provider may select the start and end date of the retrieval, and its periodicity (i.e., retrieve once, periodic retrieval according to schedule, and polling every 60 seconds). By selecting the periodic retrieval option, the data asset provider can configure the retrieval according to a schedule (or potentially multiple schedules), and the retrieval periodicity (i.e., hourly, daily, weekly, or monthly). The polling retrieval method is configured by default for ingesting data every minute. The next step of this workflow is to set the processing periodicity of the data retrieved, by selecting one of the options: (a) immediately, (b) on an hourly basis, (c) on a daily basis, or (d) on a weekly basis. Finally, the last option of this configuration is to determine the way that errors are handled. Currently there are two options available: (a) No action, and (b) Retry a specified number of retries every 30 seconds, in case that an error has occurred.

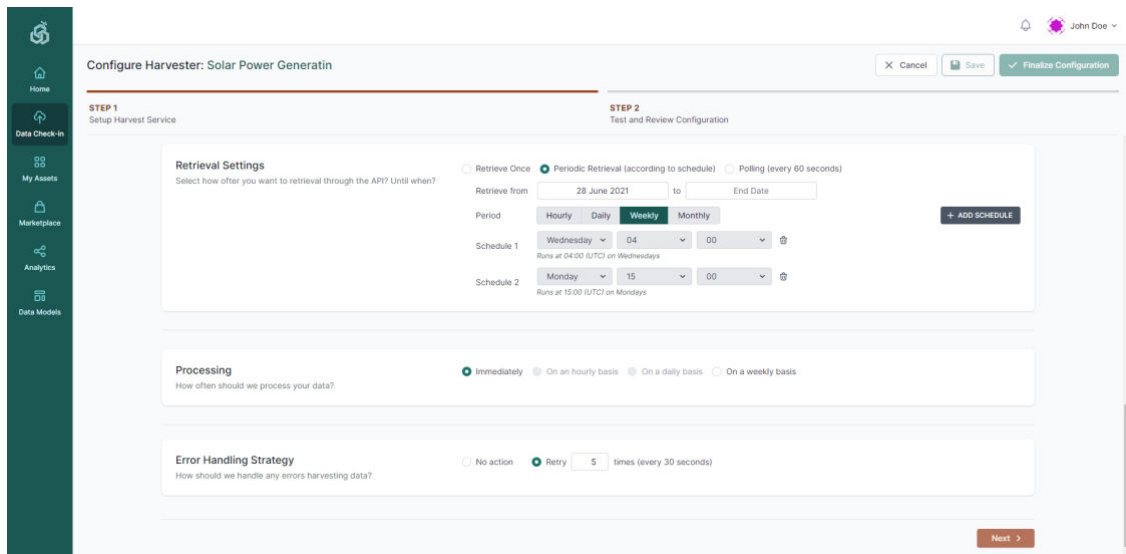


Figure 10: Data Check-in - API (external) – Setup Harvest Service (Step 1c)

By selecting the Next button, the SYNERGY platform will make a call to the API endpoint according to the configuration that has been set. If the API call is successful, the data asset provider comes across the Review step as depicted in Figure 11. During this step, the data asset provider may select the response handling method that determines whether each API response will be handled as a single object and stored as one individual record containing the concepts selected by the data asset provider, or whether each API response should be stored as multiple records that can be separated based on the selected path of the response. Additionally, the data asset provider may insert additional response data, associated with a value and its value type (e.g. static or dynamic). A static parameter added in each record/row of the response data, is a fixed value that does not change each time the API is called, while a dynamic parameter added in each record/row of the response data, contains datetime values that are updated each time the API is called.

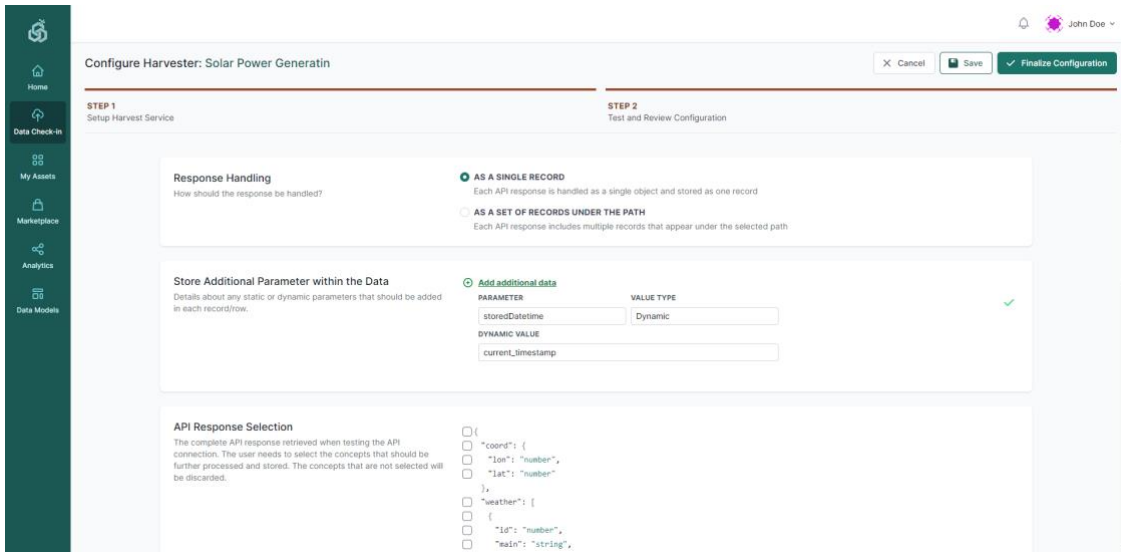


Figure 11: Data Check-in - API (external) – Review (Step 2a)

Finally, the data asset provider needs to review and select the concepts that should be further processed and stored to the SYNERGY platform, while a summary of the API’s response that will be permanently stored, will be shown accordingly at the bottom of the page as shown in Figure 12.

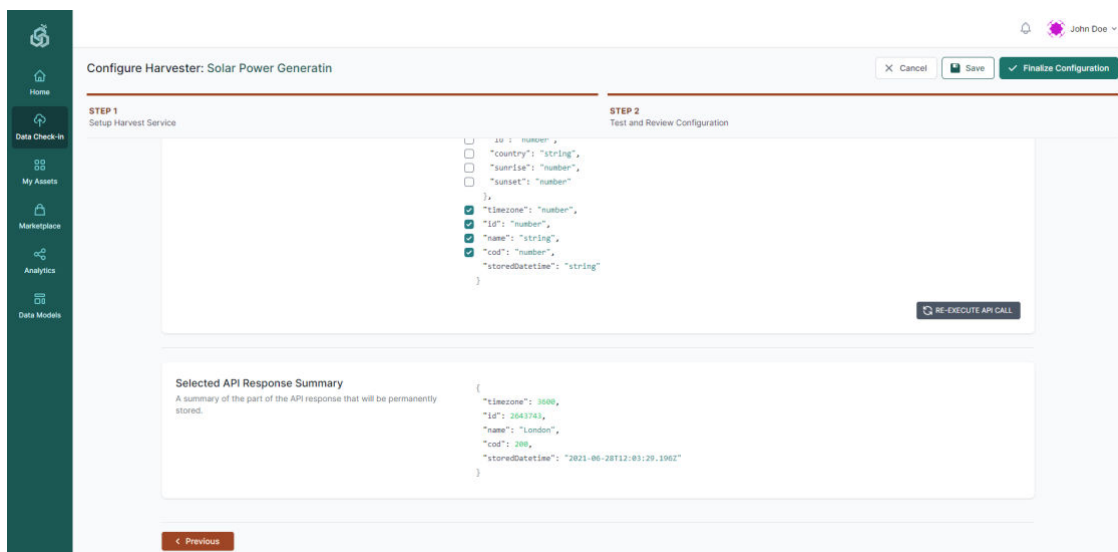


Figure 12: Data Check-in - API (external) – Review (Step 2b)

3.2.1.3 Platform’s (internal) API method

Once the data asset provider has selected the Platform’s API data harvesting method during the creation of the Check-in Job, its configuration page is shown as Figure 13 depicts. The data asset provider should select the type of the data that are to be uploaded (e.g. text, text and binary).



In particular, the data asset provider may choose whether to upload text data (e.g. JSON, or XML) or text data along with binary data (including any file format, e.g. JPG, PDF, IFC, etc) through the generated API. In addition to this, the data asset provider needs to upload a sample file containing an indicative number of entries (rows) from the “text” data that will be sent to the SYNERGY Platform’s API.

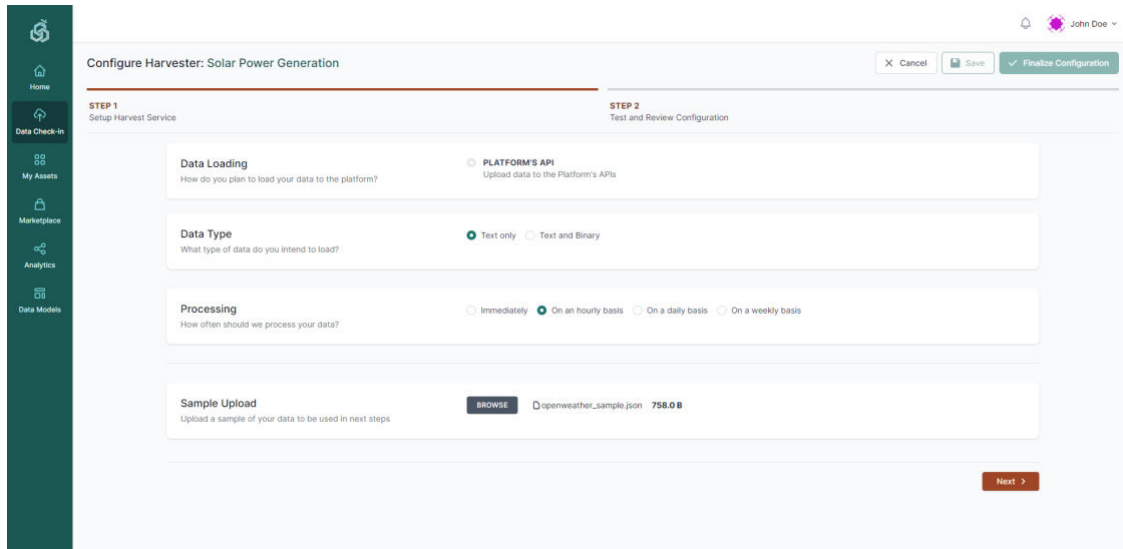


Figure 13: Data Check-in - API (internal) – Setup Harvest Service (Step 1)

During the second step (Test and Review phase) of the Platform’s API data harvesting method that is shown in Figure 14, the data asset provider is able to see an auto-generated API endpoint in the SYNERGY platform. However, to use this generated API, the application should be authenticated using an already generated access token that needs to be added into an “X-API-TOKEN” header in the API request. If the application is not authenticated, the data asset provider should generate a new access token (with “upload” scope) by selecting the “generate a new one button” link as shown in Figure 14. Instructions on how to use the POST endpoint are provided accordingly. Finally, the data asset provider may review the details of the configuration and the data sample that was uploaded in the previous step, and select Finalize to proceed to the configuration of next steps of the Data Check-in workflow (as described in Section 3.2.2).

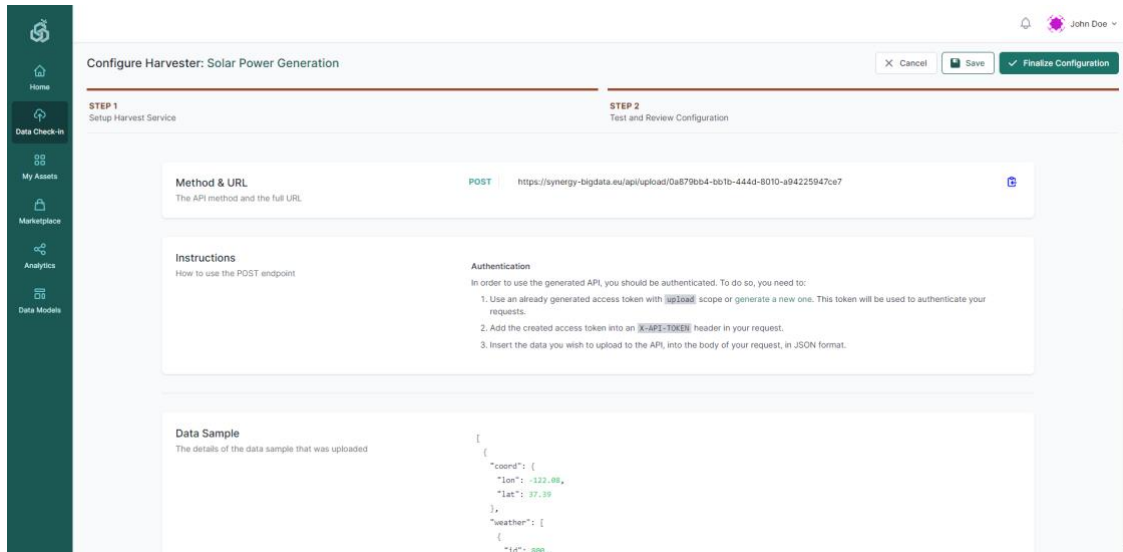


Figure 14: Data Check-in - API (internal) – Review (Step 2)

3.2.1.4 Streaming data to platform’s (internal) mechanism

Once the data asset provider has selected to harvest data through the Platform’s Kafka PubSub Mechanism, the corresponding configuration page is shown as in Figure 15.

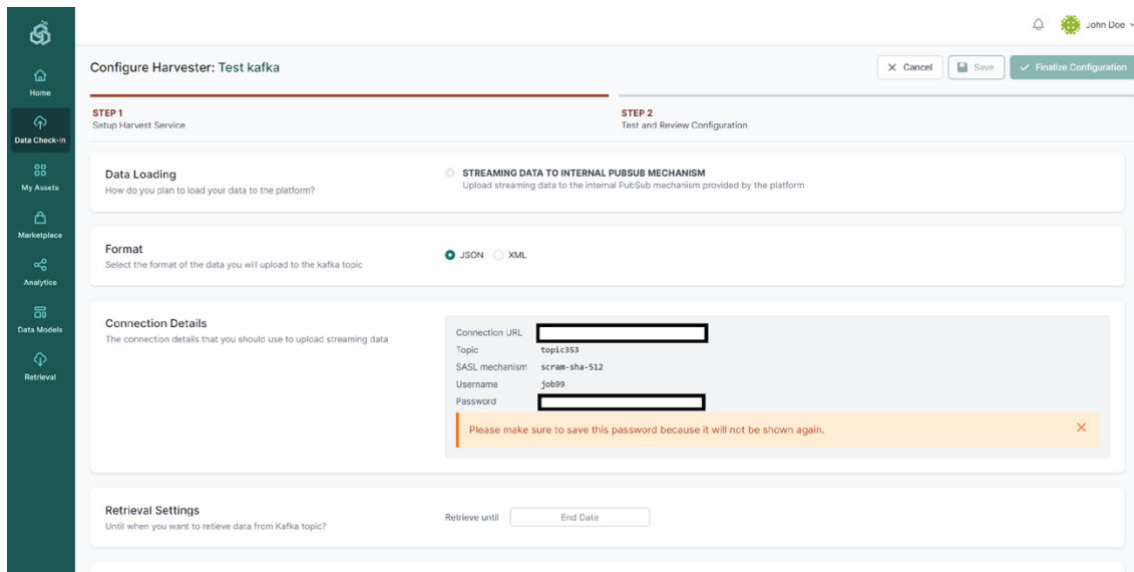


Figure 15: Data Check-in - Streaming data (internal) – Setup Harvest Service (Step 1)

During this configuration, the data asset provider should select the format of the data to be published (i.e., JSON, or XML). Then the data asset provider may view the connection details (i.e., the connection URL that refers to the platform’s Kafka PubSub mechanism, the topic name in which the data should be published, the Simple Authentication and Security Layer (SASL) mechanism that is used, and the credentials that the data asset provider should use to publish

data to the specific topic). The retrieval end date that defines the date that the streaming data retrieval has to be completed, the periodicity that defines how often the data are processed, and the error handling strategy have to be set accordingly. Finally, the data asset provider needs to upload a sample of the streaming data according to the file format selected (i.e., JSON, or XML), in order to proceed to view the sample in the next step.

During the review step, the data asset provider may finalize the harvesting configuration, after reviewing the streaming data structure (in a tree-view), and by selecting the Finalize button as depicted in Figure 16.

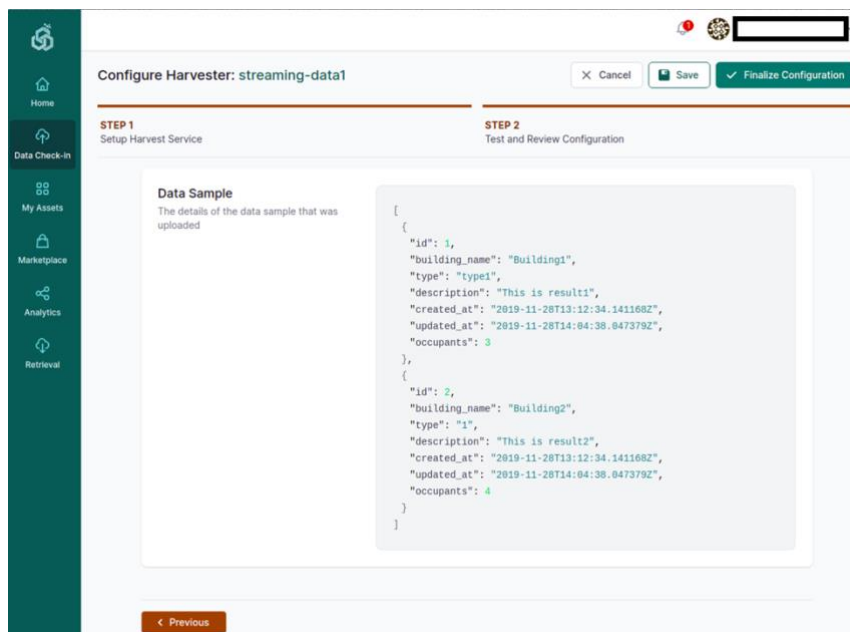


Figure 16: Data Check-in - Streaming data (internal) – Review (Step 2)

3.2.1.5 Streaming data from application’s own (external) mechanism

Once the data asset provider has selected to use a Kafka PubSub mechanism provided by an external application (owned by the data provider), the corresponding configuration page for subscribing the SYNERGY Platform to the already published streaming data, as shown in Figure 17 is revealed. Initially, the data asset provider needs to select the format of the streaming data that is published (i.e., JSON, or XML) in the specific topic, and to upload a sample of the streaming data including a few entries from the streaming data. Moreover, the data asset provider shall insert the connection URL of the external Kafka that shall be used by the SYNERGY platform, the topic name to which the data will be published, the SASL mechanism that is used,

the credentials that the SYNERGY Platform should use to access the Kafka mechanism and the group id, if applicable.

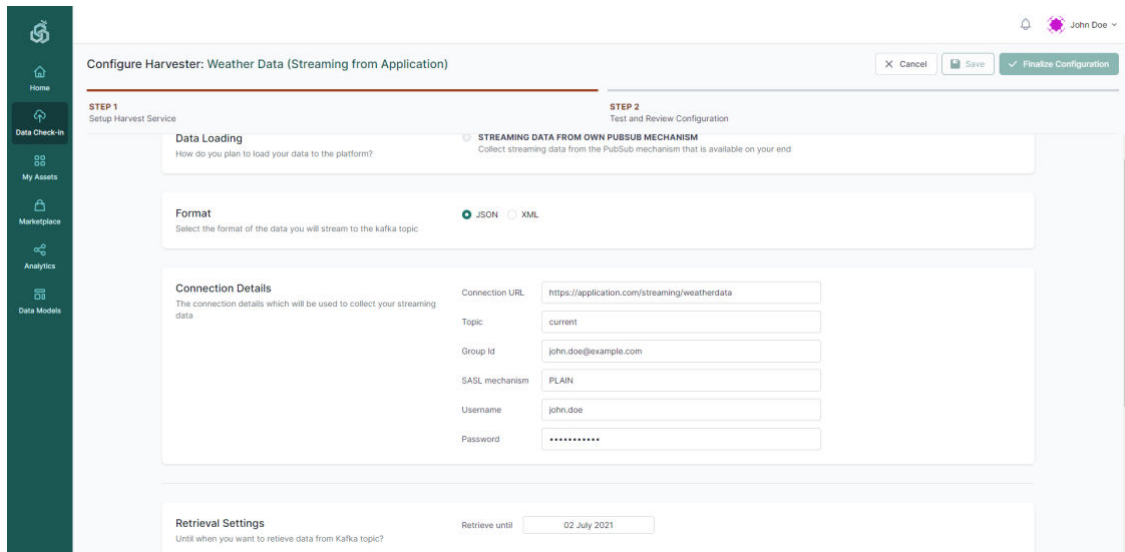


Figure 17: Data Check-in - Streaming data (external) – Setup Harvest Service (Step 1)

In the next step, the SYNERGY Platform connects to the external Kafka PubSub provider, subscribes to the topic and displays the data retrieved in order for the data asset provider to decide which fields will be kept for the next pre-processing steps of the specific data check-in job (before finalising the Harvester configuration).

3.2.2 Pre-processing Rules Definition

As mentioned in Section 3.1 which describes the creation of a new data check-in job, the data asset provider is asked to select some pre-processing rules to be applied on the data upon their ingestion in the SYNERGY Platform. The pre-processing rules that are available for selection, as shown in Figure 4, are Mapping, Cleaning, Anonymisation, and Encryption. Although mapping is not a mandatory step, it is strongly encouraged to be selected in order for the data to be mapped to the SYNERGY CIM (that has been described in the SYNERGY Deliverable D3.1). In addition, by enabling Mapping, the Cleaning, Anonymiser and Encryption steps also become available for selection, in order to solve data quality issues (e.g. corrupt or inaccurate records), and to anonymise the data to prevent an individual from being identified, respectively. It needs to be noted that the Encryption step is available only if the data asset provider has opted for on-premise execution to ensure end-to-end data security during data transfer. This section describes the workflow that a data asset provider needs to follow in order to configure these

pre-processing rules, prior to their actual execution which is described in the subsequent Section 3.3.

3.2.2.1 Mapping Configuration

The mapping configuration, provided by the Data Collection Services Bundle as described in Section 2 of D3.5 “Data Collection, Security, Governance & Management Services Bundles – Release 1.00”, is divided into three main steps namely the Mapping Info, the Mapping Configuration, and the Mapping Review and Confirmation. In the first step, the data asset provider is asked to select the standard to which the ingested data comply, if applicable, and the main category in which the data refer to, as depicted in Figure 18. Once the data asset provider selects a category, he/she needs to define the most appropriate concept that represents the data that will be uploaded.

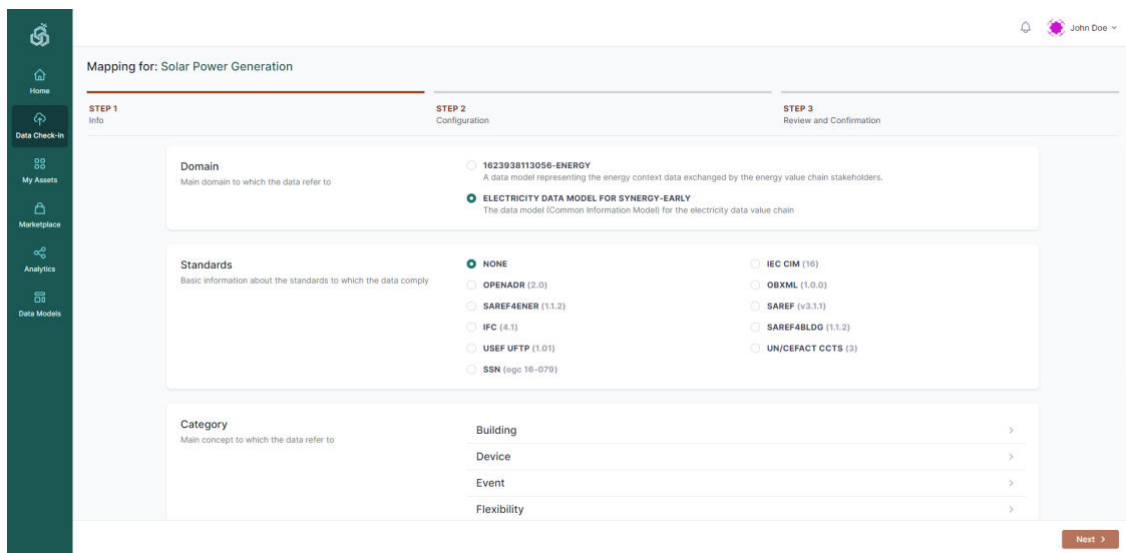


Figure 18: Data Check-in - Mapping - Provide Mapping Info (Step 1)

In the next step, the data asset provider comes across the predictions that the SYNERGY Platform has made about how the source data (from the sample provided in the Data Ingestion step) map to the SYNERGY Common Information Model (CIM) (that has been specified in the SYNERGY Deliverable D3.1) in the Mapping Playground, as depicted in Figure 19.

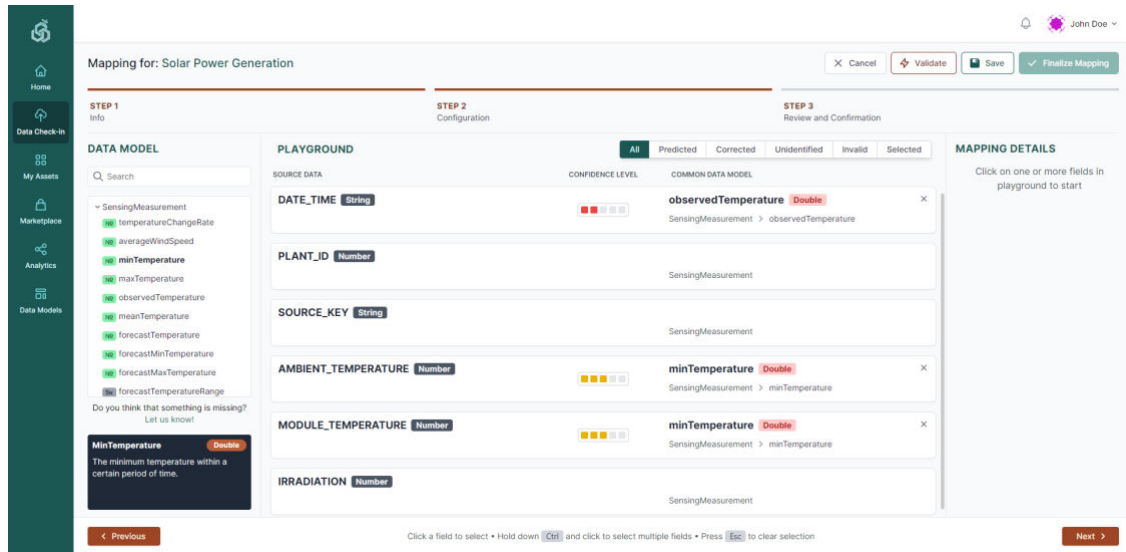


Figure 19: Data Check-in - Mapping - Configuration (Step 2a)

These predictions are accompanied by certain confidence levels denoting how much uncertainty there is for a particular field/concept to be correctly predicted and mapped to the corresponding field/concept of the common data model. In case there is no predicted mapping, or the predicted mapping is wrong, the data asset provider can search for a particular field/concept under the Data Model section located at the left side, as shown in Figure 19, while a short description of the selected field/concept appears at the bottom left side. Additionally, the data asset provider can propose a new field/concept in case the CIM has not already foreseen it, by selecting the “Let us know” link which pop-ups a new window for proposing a new field/concept to be added in the common data model, as shown in Figure 20.

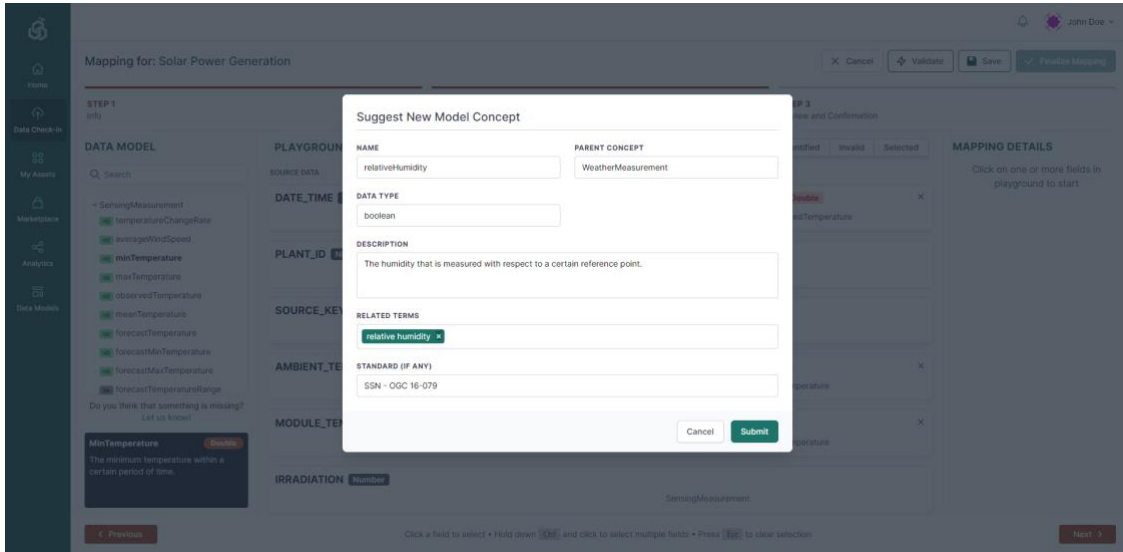


Figure 20: Data Check-in - Mapping - Suggest New Model Concept

A particular column of the source (original) data may be mapped to a specific field or concept that is listed in the Data Model section. A column in the source data can be mapped to a field of the CIM by placing the appropriate field from CIM to the corresponding source data (row in the Mapping Playground) using the drag n’ drop functionality. A column in the source data can be mapped to the fields of a related concept by selecting the corresponding row in the Mapping Playground and by selecting the related concept (along with a related prefix) from the dropdown menu that appears in the Mapping Details section located at the right side of the page, as shown in Figure 21. In cases the related concept can be customised (according to the CIM provisions), the data asset provider can provide his/her own prefix along with its definition.

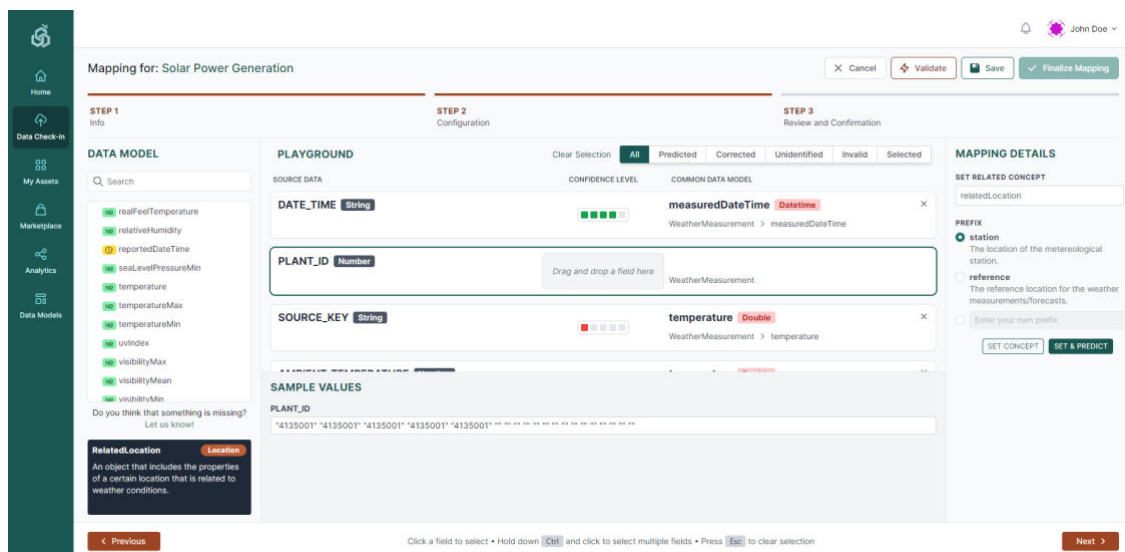


Figure 21: Data Check-in - Mapping – Configuration (Step 2b)

By selecting the Set Concept, the corresponding list of all the fields included in the related concept is revealed in the left Data Model section, allowing the data asset provider to map the input data (row in the Mapping Playground) to a particular field that is included in the related concept selected. This allows the connection of fields in the source data with other fields that are included in the selected related concept of the base concept that was selected in Step 1. The same procedure should be followed when the data asset provider selects the Set Concept & Predict, although in this case, the prediction service is executed once more to predict mappings between the input data and the selected related concept.

In the right section, the Transformation Details are shown providing more information including the transformations rules that are to be applied on the different fields depending on their data type and the CIM provisions. For datetime fields, the desired datetime format and the applicable time-zone (if enabled) should be defined as shown in Figure 22. It needs to be noted that during the mapping step, data type casting according to the CIM provisions is also performed (e.g. from integer to double, from string to datetime, etc.).

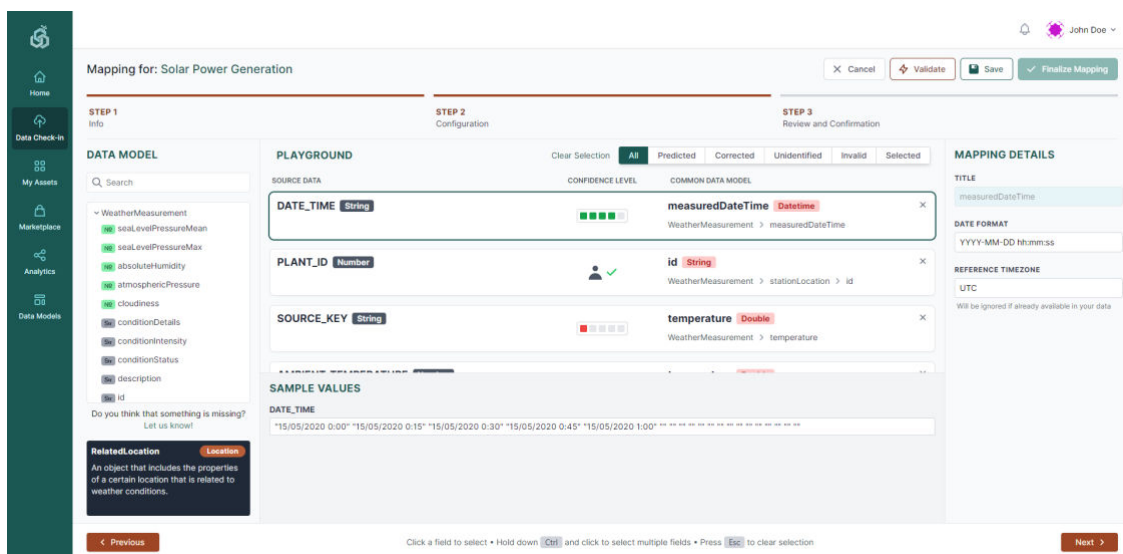


Figure 22: Data Check-in - Mapping Configuration - Datetime Field Transformation

In the Mapping Playground section, the data asset provider may filter the view of the concepts by different categories (i.e., Predicted Mappings, Unidentified Mappings, Corrected Mappings, Invalid Mappings, and Selected Mappings). Once the data asset provider is satisfied with the mapping configuration, by clicking the Next button, the Mapping Review and Confirmation page

is loaded as depicted in Figure 23. During this step, the data asset provider may view a summary of the concepts that are mapped to the SYNERGY CIM (described in D3.1 “SYNERGY Common Information Model”) and will be transformed according to the configuration provided during the previous step. Additionally, the data asset provider may view more details regarding the mapping and transformation for the source data (per field) by selecting a particular row as shown in Figure 23. In addition, users are able to view the Unidentified Concepts at the bottom of the page, where the title and data type of the concepts that are not mapped to the CIM, and thus they will be excluded from the mapped data that will be uploaded in the SYNERGY Platform and proceed to further processing.

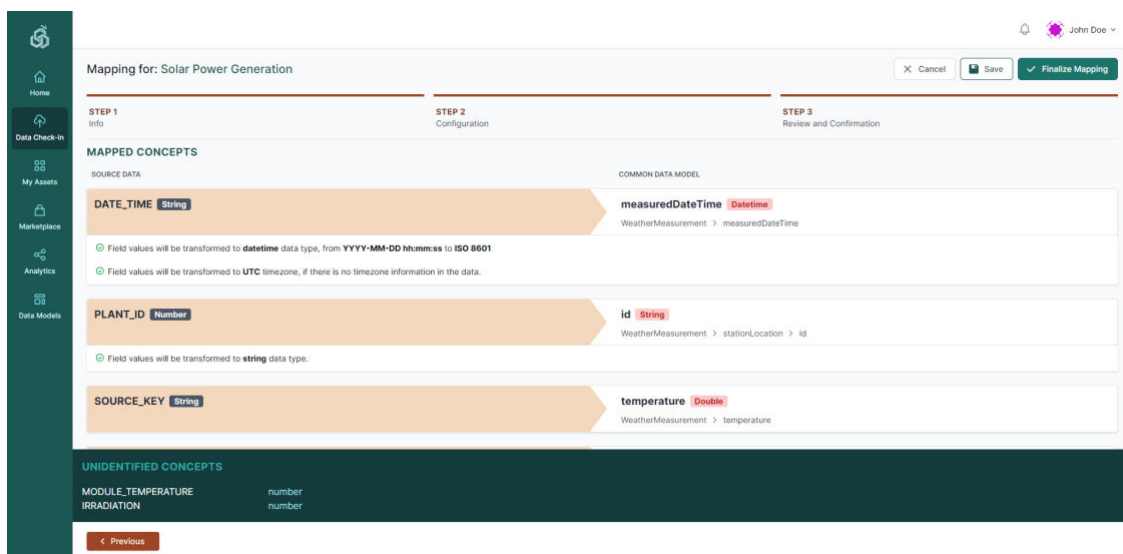


Figure 23: Data Check-in - Finalize Mapping (Step 3)

3.2.2.2 Cleaning Configuration

If the cleaning step has been selected during the creation of the data check-in job, as shown in Figure 4, the corresponding workflow for manipulating and cleaning the data ingested into the SYNERGY Platform needs to be configured. This functionality aims at providing to the SYNERGY Platform, accurate, complete, and consistent data that could be used by the specific organization that owns the data asset, but also the overall electricity value chain that can potentially acquire it. The workflow on this functionality involves several features offered by the Cleaning Service described in Section 2 of D3.5 “Data Collection, Security, Governance & Management Services Bundles - Release 1.00”. In particular, the data cleaning workflow involves the definition of data cleaning rules depending on the data type of each field in the dataset, in order to eventually store a high-quality dataset. The cleaning rules are divided into validation

options and corrective actions. The former involves the definition of allowed value ranges, uniqueness constraints, mandatory constraints, regular expression patterns, and outliers identification, while the latter involves dropping unnecessary columns/entries and replacing values when needed. The first step of the data cleaning workflow allows the data asset provider to select the fields/columns to which he/she will define the cleaning rules and constraints, as shown in Figure 24.

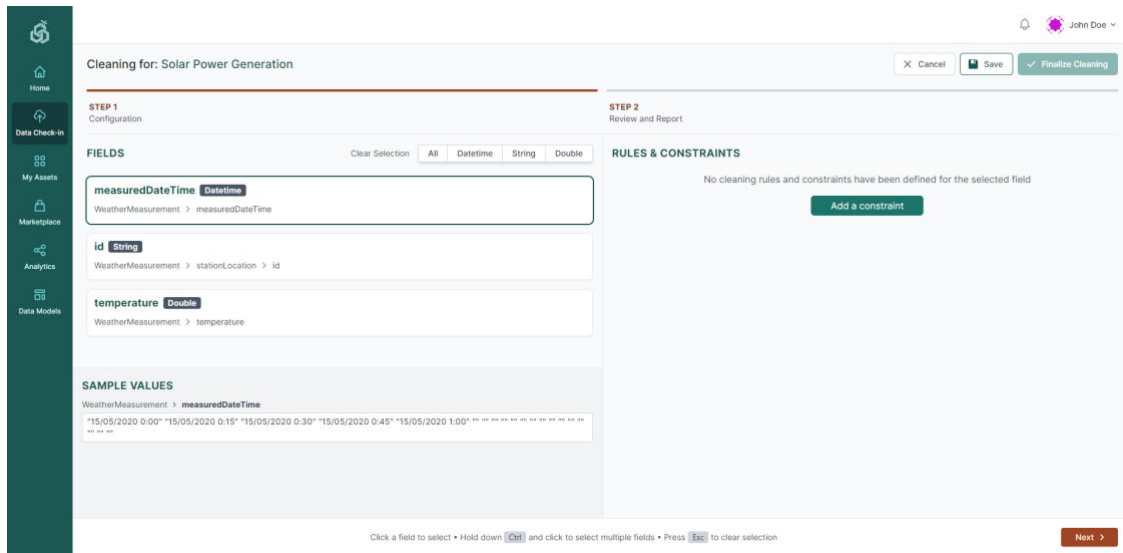


Figure 24: Data Check-in - Cleaning Configuration – Select Fields for Cleaning (Step 1a)

It needs to be noted that the data asset provider is able to select multiple columns based on the same data type, in order to define rules that are applicable to more than one fields. As shown in Figure 25, the data asset provider may define different constraints (e.g. mandatory constraint where field values must not be null, regular expression pattern constraint where field values must have an exact match with a particular regular expression, and unique constraint where

field values must be unique). Upon this selection, he/she may define an outlier rule (e.g. drop, or replace with a particular default value) in case a value is considered as outlier.

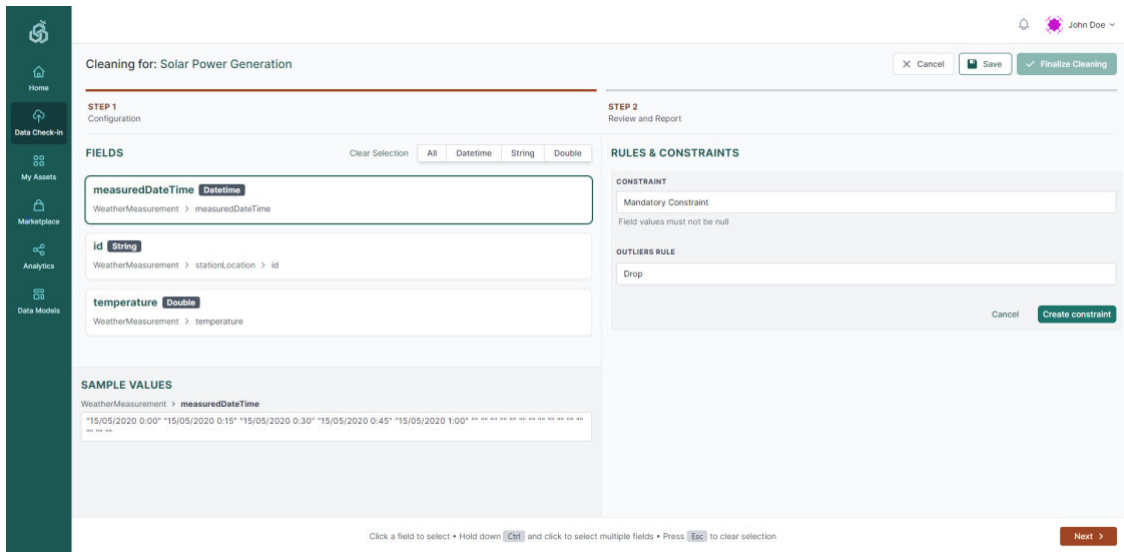


Figure 25: Data Check-in - Cleaning Configuration - Set Rules & Constraints (Step 1b)

Each time a cleaning rule and constraint is added, the data asset provider can view what exactly will be applied in the data in simple language. In addition, the data asset provider may add a new rule and constraint, edit or remove an existing rule or constraint, or even change the order of the rules or constraints that are to be applied as shown in the Rules & Constraints section at the right part of Figure 26.

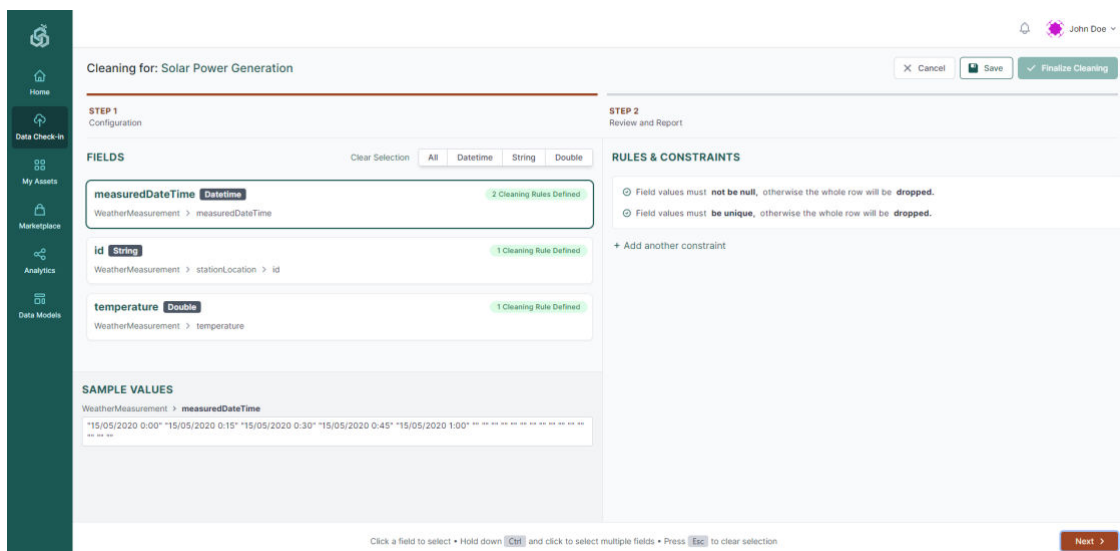


Figure 26: Data Check-in - Cleaning Configuration - View Rules (Step 1c)

Once the data asset provider is satisfied with the cleaning configuration, by selecting the Next button, the final step of reviewing the cleaning rules and constraints will appear as shown in Figure 27.

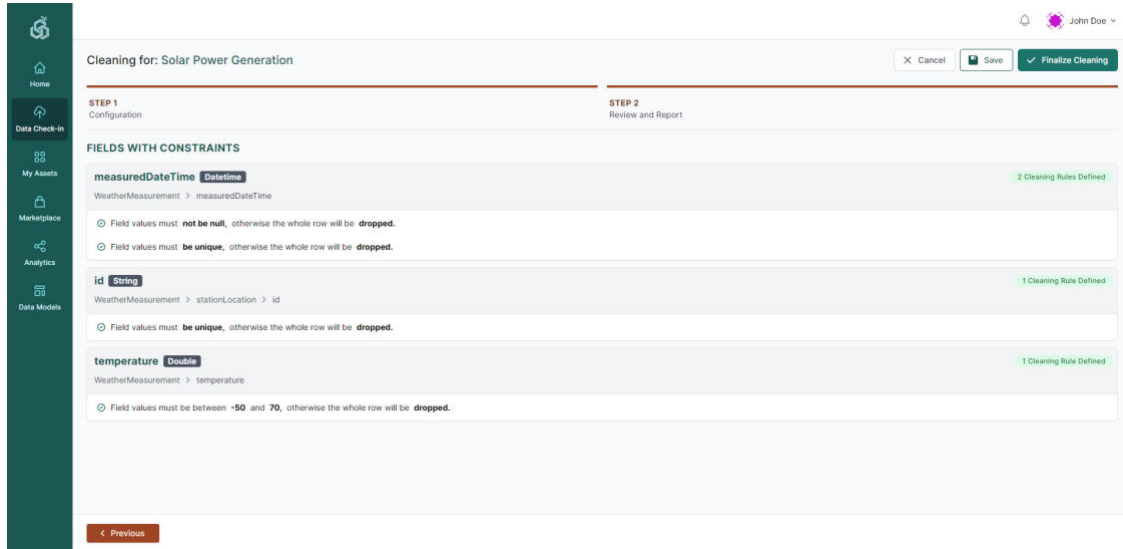


Figure 27: Data Check-in - Finalize Cleaning (Step 2)

3.2.2.3 Anonymisation Configuration

Another optional pre-processing step of the data check-in job configuration is the Anonymisation step, offered by the Data Collection Services Bundle as described in Section 2 of D3.5 “Data Collection, Security, Governance & Management Services Bundles - Release 1.00”. In particular, if the anonymisation step is enabled in the data check-in job configuration, the execution of the Anonymisation Service, which is described in Section 4.2.1 of D2.7 “SYNERGY Framework Architecture including functional, technical and communication specifications v2”, will be triggered, and the data asset provider shall define the data anonymisation rules to be executed as shown in Figure 28. During this step, a data asset provider defines which of the fields that have been mapped to the SYNERGY CIM during the Mapping step are sensitive, quasi-identifiers and identifiers since by default all fields are considered as insensitive. Sensitive data (that reveal private information such as genetic data, health data, ethnic origin, etc) should be protected by an anonymisation algorithm that ensures that an individual cannot be identified. If the data asset provider selects identifier (i.e. field that can be directly used to identify an individual) as the anonymisation type, then the particular column will be dropped from the dataset. Otherwise, if the data asset provider selects quasi-identifier (i.e. field that is not an

identifier itself, but combined with other quasi-identifiers may uniquely identify an individual), the generalization method needs to be defined depending on the field type.

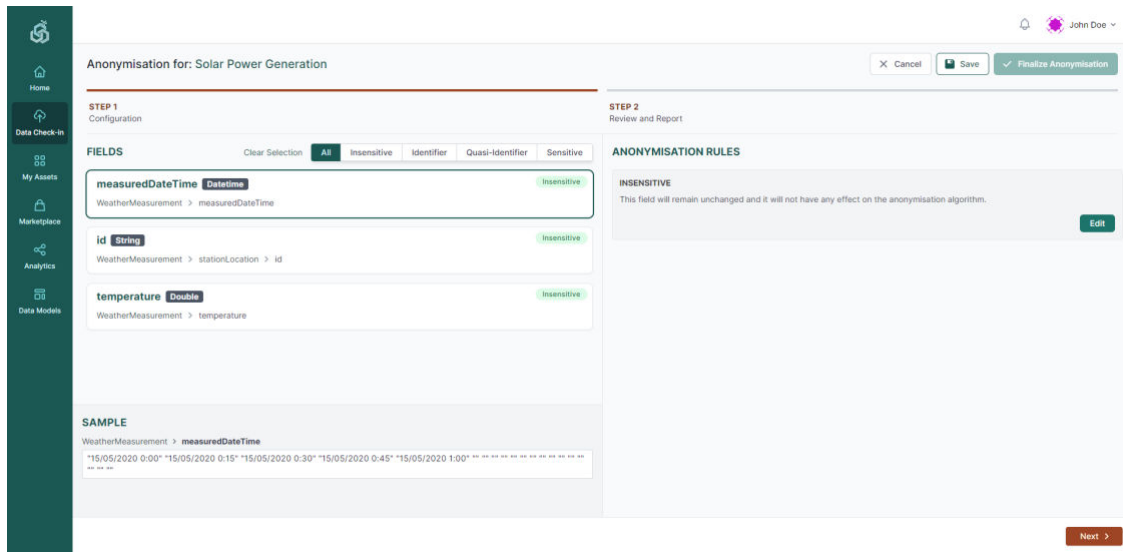


Figure 28: Data Check-in - Anonymisation Configuration – Select Fields for Anonymisation (Step 1a)

To select one of the anonymisation rules, the data asset provider needs to select the field from the Fields section, and then select the appropriate rule by selecting the Edit button from the Anonymisation Rules section at the right section as shown in Figure 29. Depending on the data type and the anonymisation type that the data asset provider selected for a field, different anonymisation methods become available providing different configuration options. Since the cleaning step is not compulsory and the null values may not be handled till this step, the data asset providers may need to select the way that the anonymisation step will handle the null values (i.e., keep, or replace with a value).

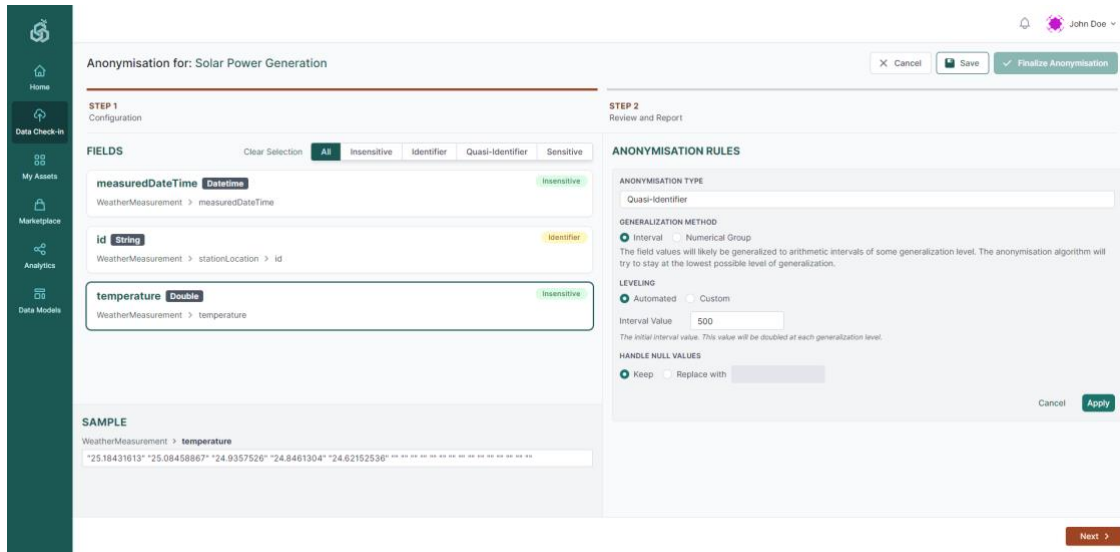


Figure 29: Data Check-in - Anonymisation Configuration - Set Anonymisation Rules (Step 1b)

As soon as the data asset provider is satisfied with the definition that he/she has provided for a field, he/she may select the Apply button in order for the Anonymisation Rules Preview section to appear as shown in Figure 30. In this section, the different anonymisation levels and examples are provided in order to make clear to the data asset provider what exactly will happen to the data.

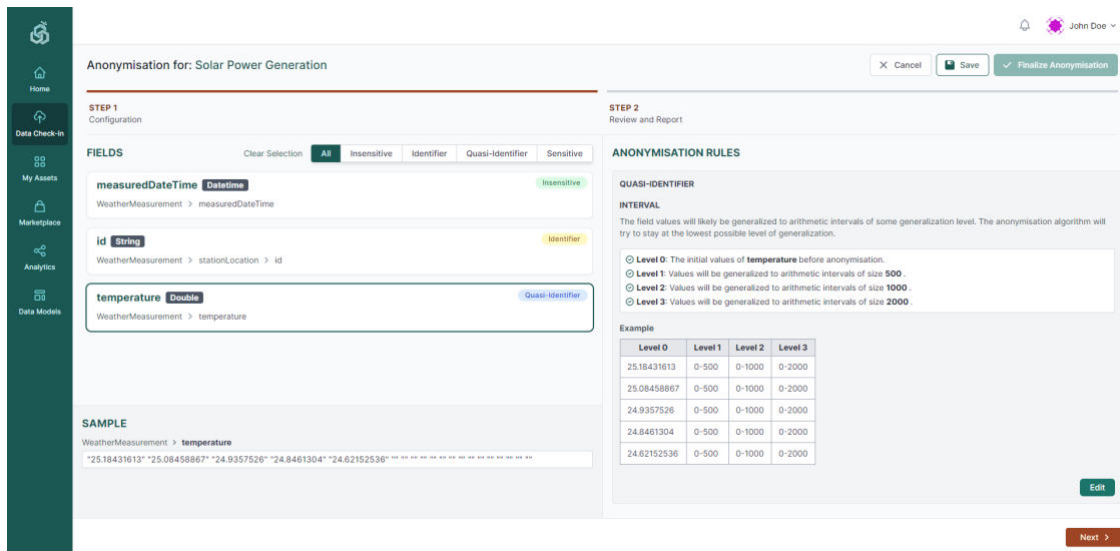


Figure 30: Data Check-in - Anonymisation Configuration - View Anonymisation Rules (Step 1c)

Once the data asset provider has defined anonymisation rules for the fields of interest, he/she may proceed to the next step where the full set of anonymisation rules can be reviewed per anonymisation type, as displayed in Figure 31. The data asset provider may also define the

acceptable information loss threshold and specific parameters of the selected anonymisation algorithm (e.g. for k-anonymity, the k-value needs to be filled in). It needs to be noted that if the desired anonymisation is not achieved or the resulting data loss is above the acceptable information loss threshold, the SYNERGY Platform will impose failure of the anonymisation step on purpose at execution time.

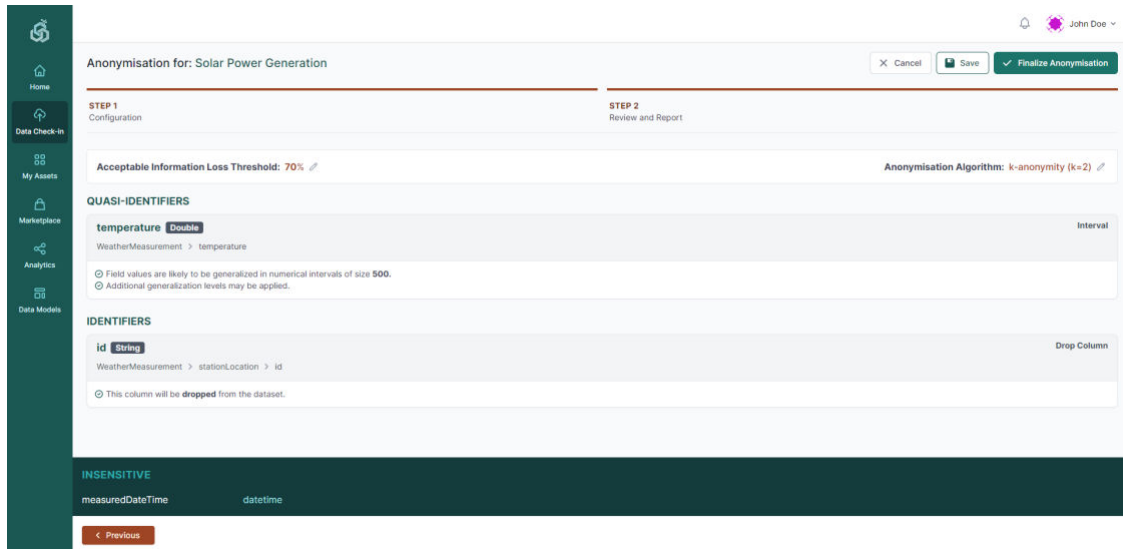


Figure 31: Data Check-in - Finalize Anonymisation (Step 2)

3.2.2.4 Encryption Configuration

The data encryption functionality is available to data asset providers that require end-to-end security in the SYNERGY Platform and the On-Premise Environment (that has been installed locally) to eliminate the risk of unauthorized data access or leakage during the data transfer to the cloud. As Section 4.2.2 of D2.7 describes, data asset providers are able to (optionally) set encryption parameters to be applied on the whole dataset. It needs to be noted that, during the creation of the data check-in job, the data asset provider is asked whether to include the data encryption in the pre-processing steps (as in Figure 4).

As depicted in Figure 32, the data asset provider is able to select the concepts of the dataset whose values are to be indexed to facilitate search (as presented in section 4.1).

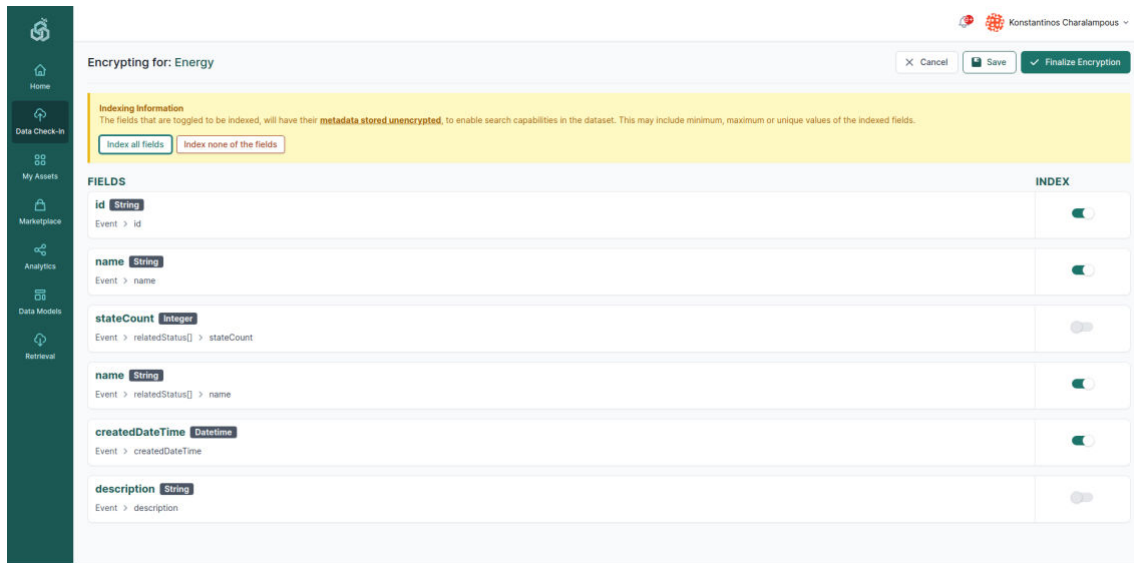


Figure 32: Data Check-in - Encryption Configuration

3.2.3 Define New Data Asset Profile

The last step of the data check-in job configuration is to store the final processed data asset in the SYNERGY Platform. Hence the data asset provider is asked to provide a title and a short description for the processed data asset that is to be stored in the SYNERGY Platform, as shown in Figure 33.

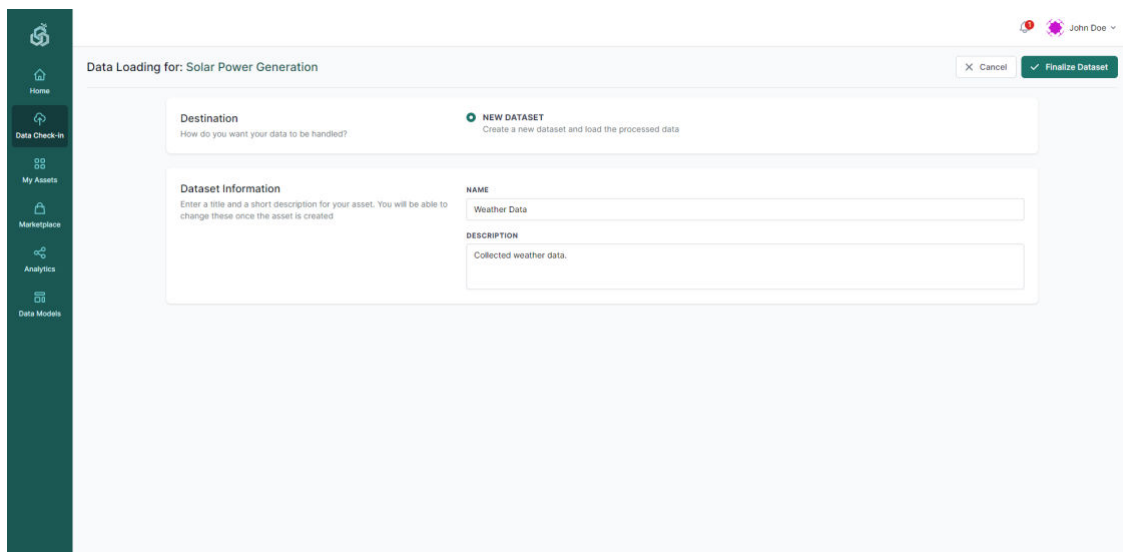


Figure 33: Create a new Data Asset Profile

However, a more detailed data asset profile needs to be defined by the data provider, according to the SYNERGY Metadata Schema. Thus, the data asset provider needs to navigate to My Assets -> Datasets on the main navigation bar. As depicted in Figure 34, the data asset provider may change the title and the description of the stored data asset, and insert additional information such as tags, distribution details, extent details, licensing information, and pricing details, according to the SYNERGY metadata schema. Depending on the access level of the data asset, the data asset provider is requested to define the applicable access policies.

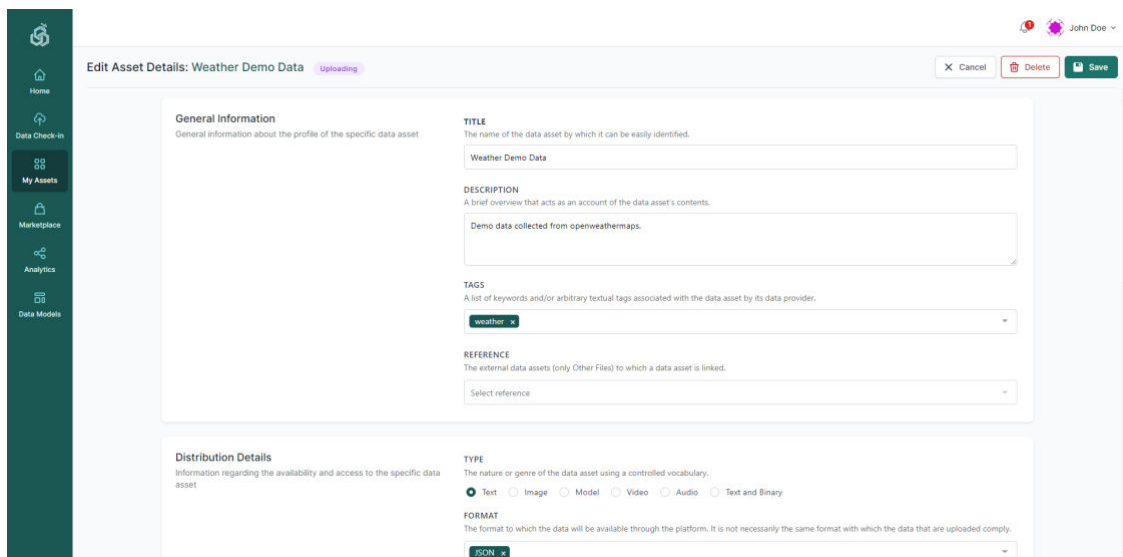


Figure 34: Define a new Data Asset Profile

In particular, the data asset provider needs to add tags (i.e., list of keywords, or arbitrary textual tags) that are associated with the data asset, as well as any potential reference to external data asset (already stored in the SYNERGY Platform) to which a data asset is linked. The type which describes the nature of dataset, the format of the data asset to which the data will be available, and the language of the data asset, need to be defined appropriately by selecting an item from the corresponding drop-down lists. Details regarding the temporal coverage, spatial coverage, temporal resolution, and spatial resolution units need to be defined as well by selecting the appropriate values for these units from the corresponding drop-down lists. It needs to be noted that for cases that the temporal or spatial coverage value cannot be specified in advance, the platform allows data asset providers to select temporal and spatial coverage based on the actual data that are uploaded, by selecting the “Calculated based on data” option.

Then, the data asset provider needs to set the access and licensing information which corresponds to the desired visibility of a data asset and the applicable access policies. Thus, the

data asset provider needs to select the access level from three different options: (a) Public, allowing the access of the data asset to any organisation without requiring any access policies to be satisfied or the existence of a data contract; (b) Private, allowing other organisations to view the data asset in the SYNERGY Marketplace if the access policies are satisfied and access it if there is an active contract; and (c) Confidential, denying access of the data asset to external organisations as it is only intended for use within the organisation that owns it. In case that the access level is set to Confidential, there are no available options regarding licensing and its associated terms under which the data asset is made available, since it will be visible only to the data asset owner.

In contrast, if the access level is set to Public or Private, the data asset provider needs to fill the corresponding licensing information, as shown in Figure 35 and Figure 36, respectively. In particular, the name of the copyright owner and the data license stating the legal terms and giving the official permission to the data asset should be selected (or provided in case of custom licenses). If the data asset provider selects an already existing well-defined data license, then the licensing details are automatically filled, otherwise the data asset provider should select Custom to fill the licensing details according to his/her needs, as shown in Figure 36. Public access level allows the data asset profile to be available to all and thus the pricing details and access policy sections are disabled since the asset is accessible by everyone.

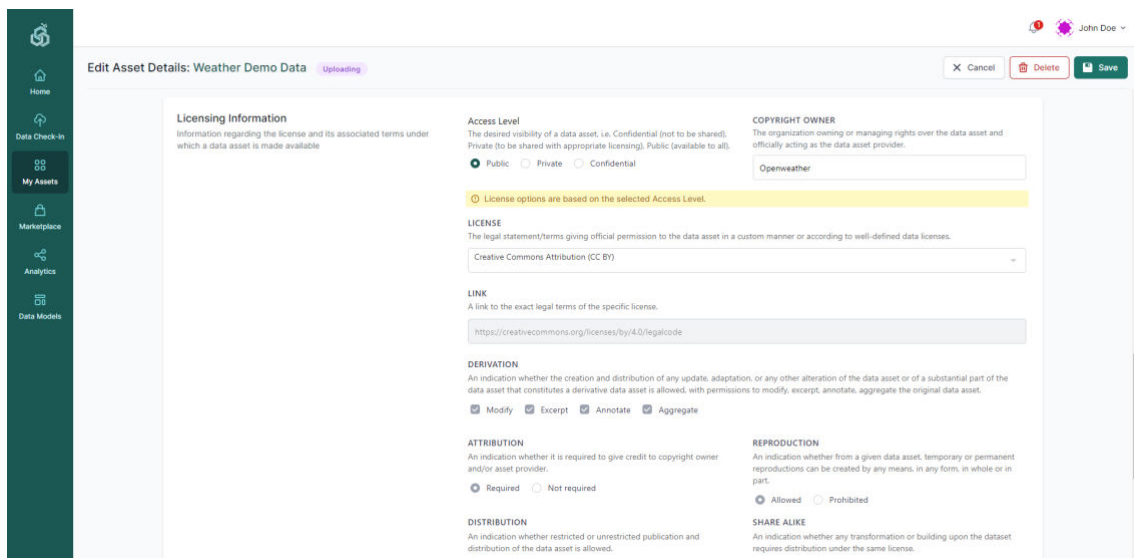


Figure 35: Define a New Data Asset Profile - Access Level - Public

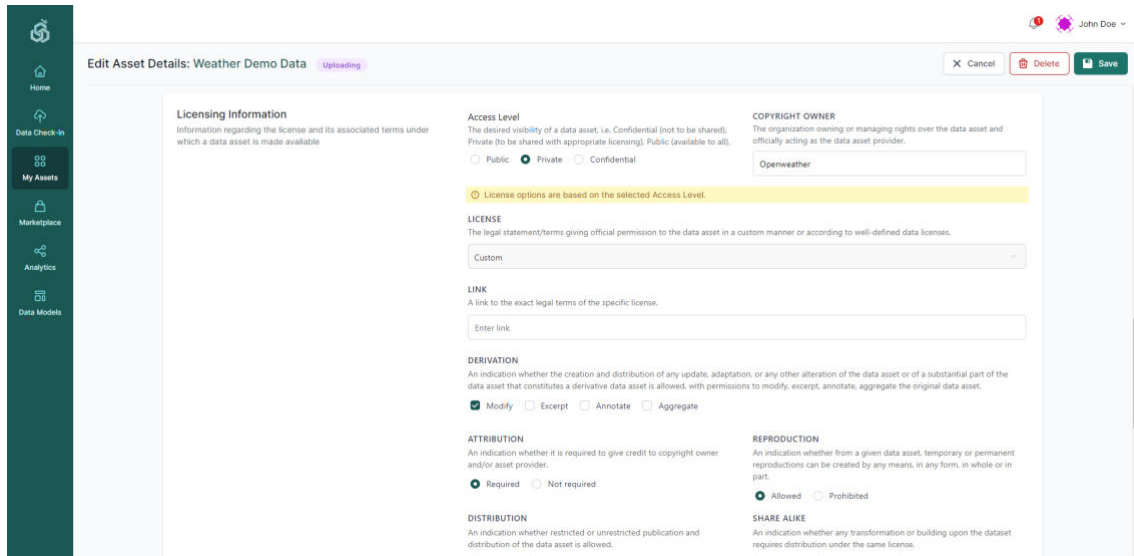


Figure 36: Define a New Data Asset Profile - Access Level - Private

By selecting Private access level, the data asset provider needs to add all the corresponding licensing information as mentioned before, along with pricing details regarding the payment and acquisition of the data asset, and the access policy defining who has access to the data asset or not. In particular, the access policies need to be defined dictating whether certain organizations (or everyone) should be allowed or denied to view the data asset in the Marketplace (note: viewing a data asset in the Marketplace only means that the data asset consumer is eligible to acquire it – a contract needs to be put in place to get actual access to it as described in Sections 4.2-4.3), by adding exceptions on certain organization or user parameters using logical conditions, e.g. on the organization type, as depicted in Figure 37.

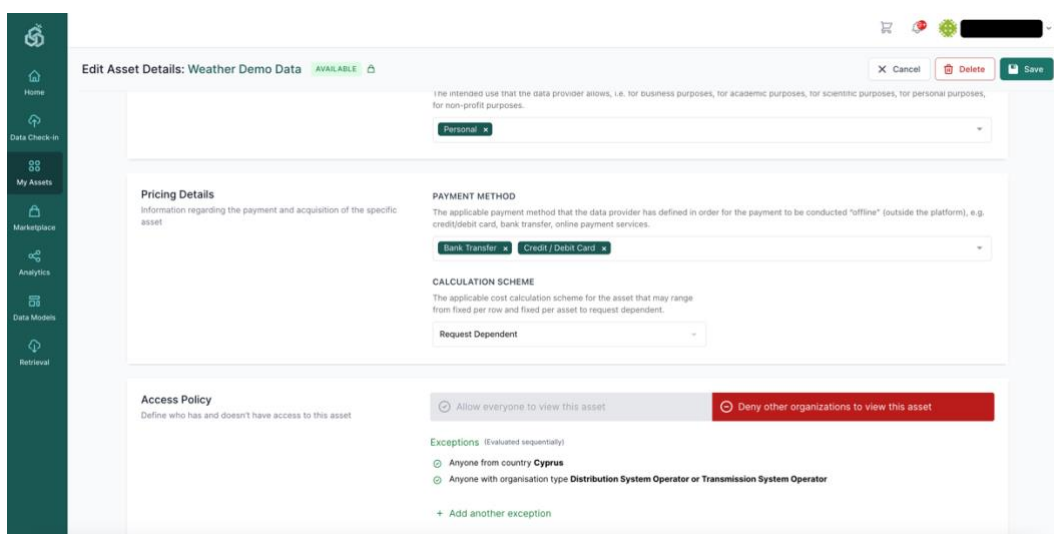


Figure 37: Define a New Data Asset – Private Access – Deny Everyone with Exceptions

By selecting Save, the data asset profile is finalized, and the data asset provider may navigate to the overview, the license details and the data structure of the data asset profile that has been created and stored. The data asset profile overview, the license details and the data structure pages are indicatively depicted in Figure 38 and 39.

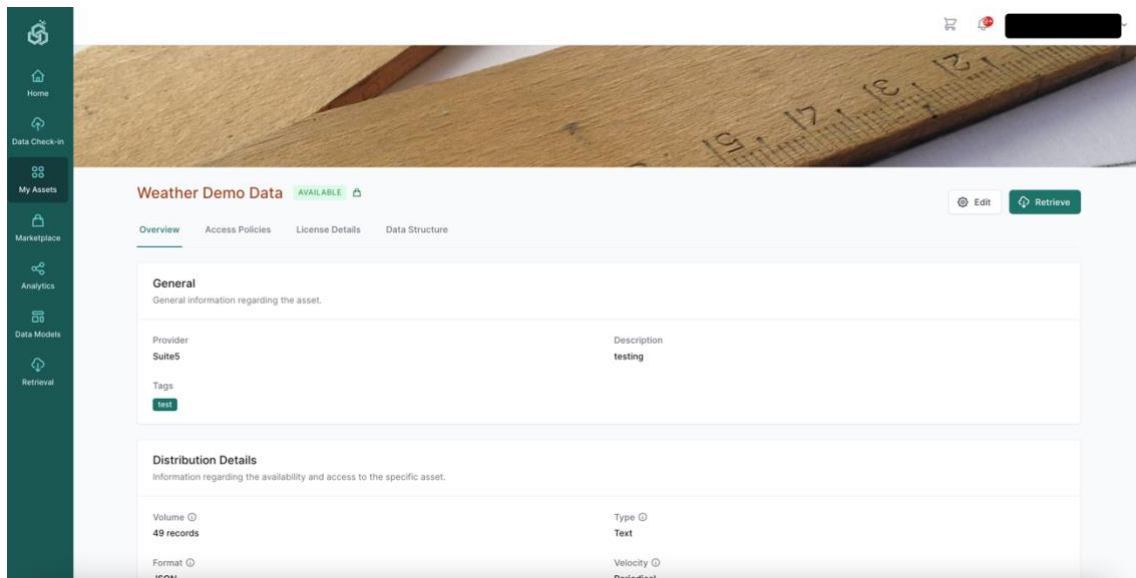


Figure 38: Define a new Data Asset Profile – Overview (visible to all)

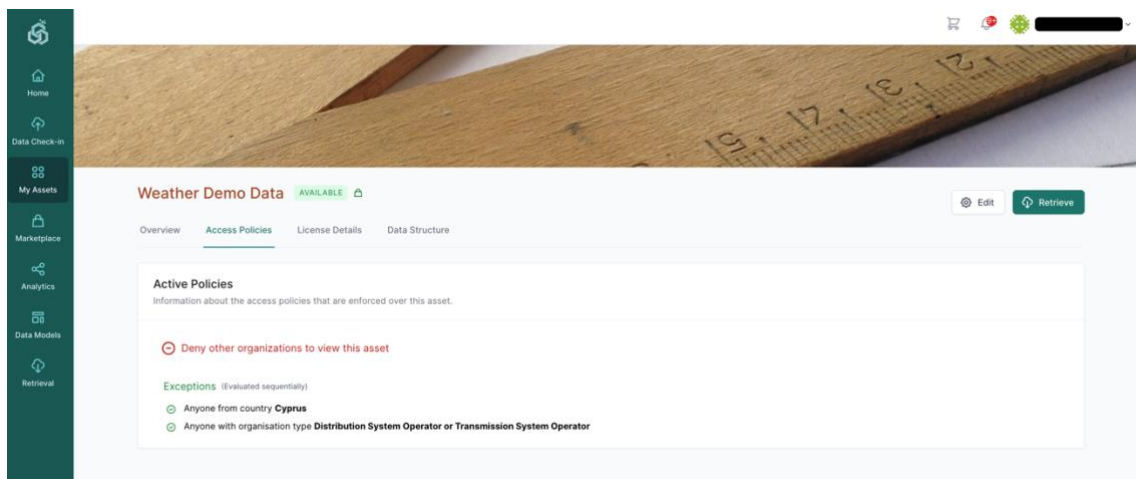


Figure 39: Define a new Data Asset Profile – Access Policies (visible only to the respective data asset provider)

Obviously, the data asset provider is able to edit the metadata (with the exception of the Data Structure) at any point through the My Assets->Datasets menu.

3.3 Execute a Data Check-in Job

3.3.1 Pre-processing Rules Execution

As mentioned already, the pre-processing steps configuration (at design time) has a distinct separation from the actual execution of each pre-processing step. In this section, a short description regarding the execution of each pre-processing step is provided accordingly. Note that, all the pre-processing steps (i.e., Mapping, Cleaning, Anonymisation, and Encryption) can be only executed if each pre-processing configuration has been finalised.

3.3.1.1 Mapping Execution

If the Mapping step has been successfully executed, the data asset provider may view what transformations happened on the data per field in the source data, as shown in Figure 40.

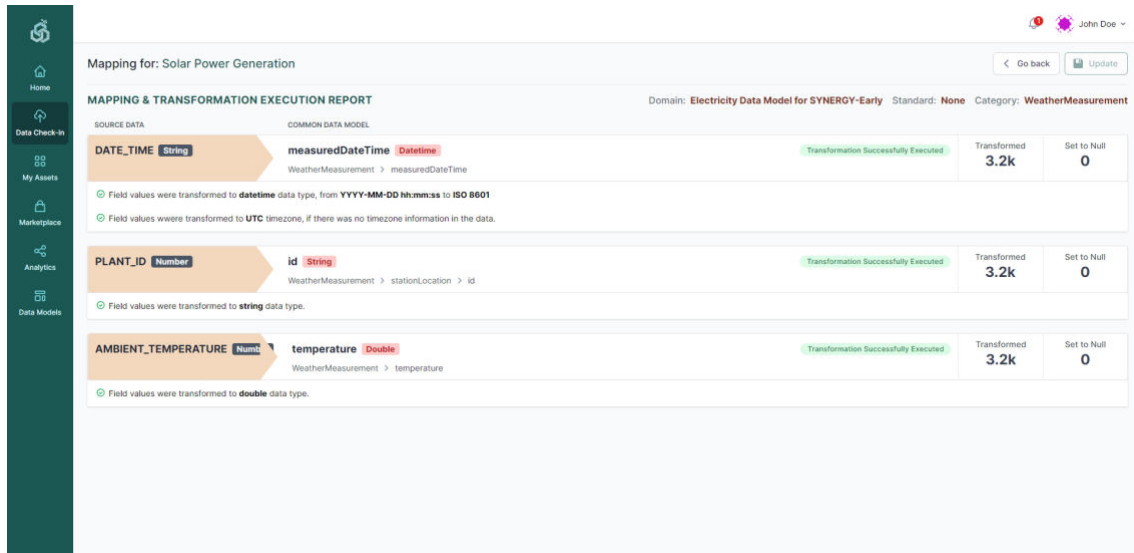


Figure 40: Mapping Execution - Successful

Otherwise, if the Mapping step failed due to wrong transformation rule definitions, the data asset provider may view the failed transformations that have been defined (as shown in Figure 41) and make the appropriate corrections in order to restart the specific step (if it has never been successfully executed). It needs to be noted that in the case of recurring jobs, the Mapping step may have run successfully for n times and failed in the n+1 time, which is why exact number of rows/values are provided per execution.

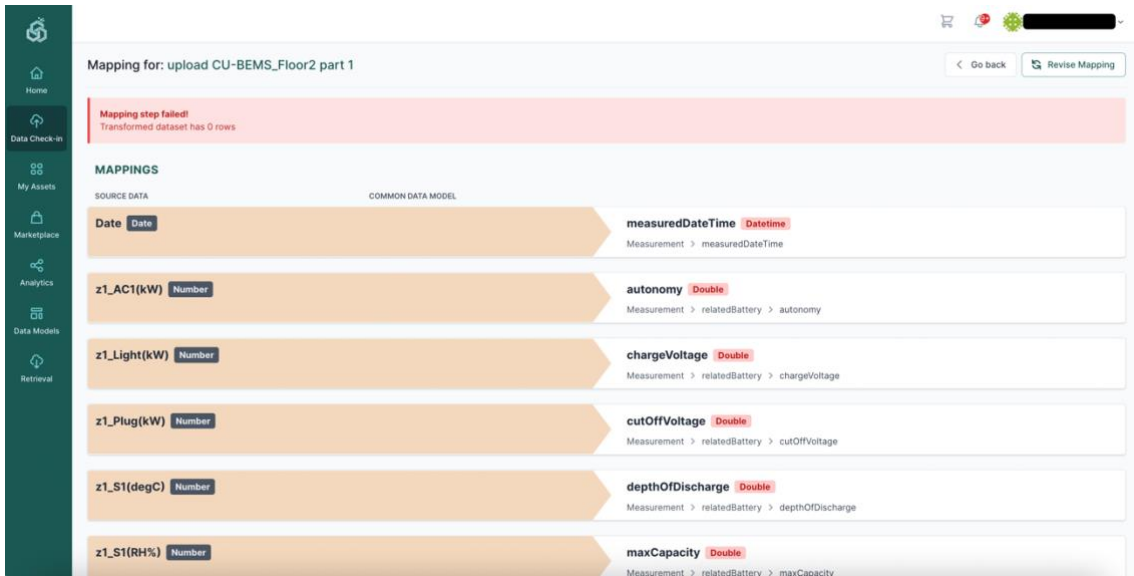


Figure 41: Mapping Execution – Failed with option to Revise step

3.3.1.2 Cleaning Execution

Once the data provider completes the Cleaning configuration and it is executed, a pop-up appears notifying the user that the Cleaning step has been executed. An indicative example of a successful execution of the Cleaning rule is shown in Figure 42. In this step, the data asset provider may view what cleaning constraints were met and resulted into transformations per field in the source data.

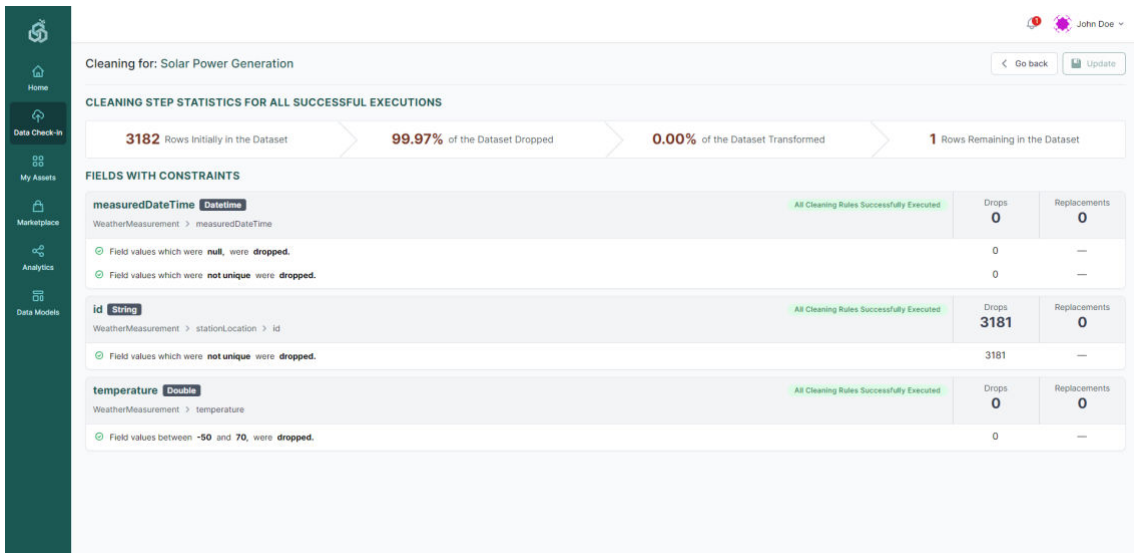


Figure 42: Cleaning Execution - Successful

However, in the case where the field values of a column do not follow the defined pattern, as configured in the Cleaning configuration rules, a failed Cleaning execution report page will appear, notifying the data asset provider to view the errors that have occurred, in order to revise the cleaning rules and constraints that have been defined.

3.3.1.3 Anonymisation Execution

If the Anonymisation step has been executed and the achieved information loss is less than the acceptable information loss threshold that has been defined during the configuration, the Anonymisation step is considered as successful as Figure 43 depicts.

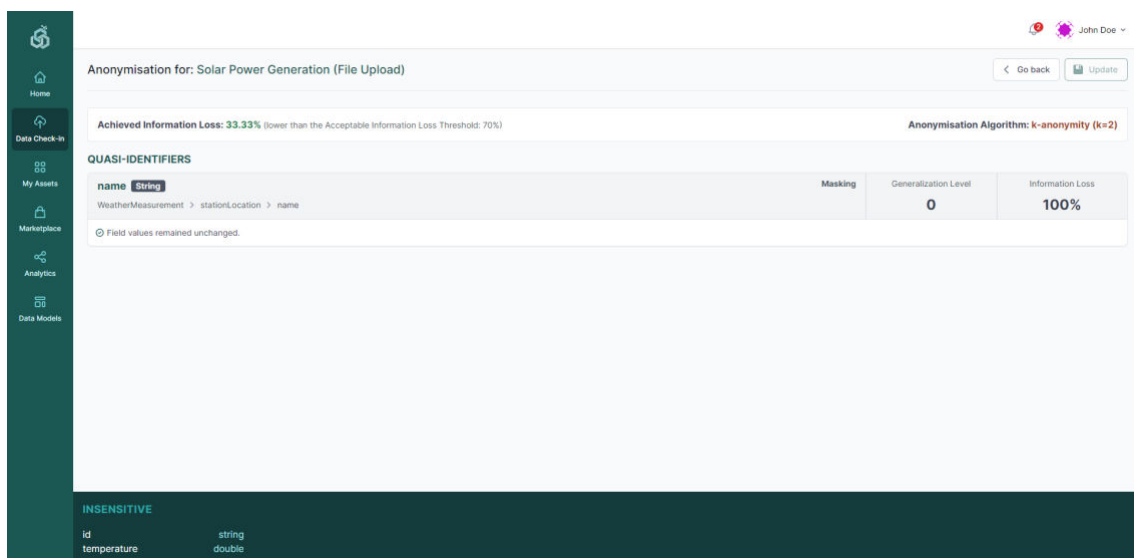


Figure 43: Anonymisation Execution - Successful

On the contrary, if the achieved information loss is higher than the acceptable information loss threshold that had been defined during the configuration or if the desired level of anonymisation was not achieved, the Anonymisation step is considered as failed.

3.3.1.4 Encryption Execution

Similarly to the rest of the pre-processing steps, the data asset provider is able to view the report of the encryption execution, displaying whether the encryption step was executed and applied successfully or not (according to the configuration provided), as depicted in Figure 44.

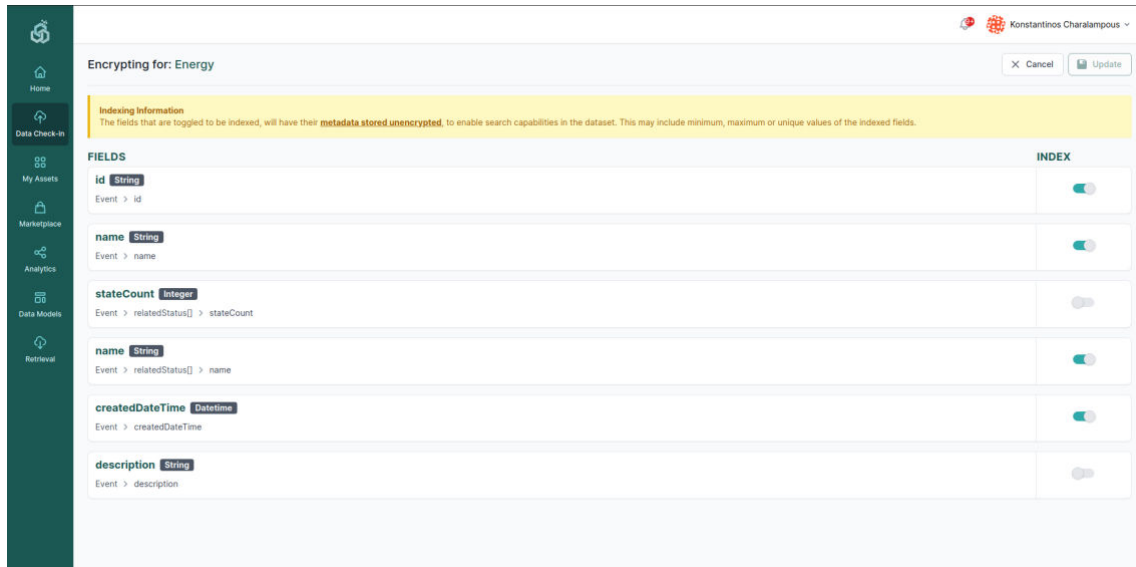


Figure 44: Encryption Execution

3.4 Upload Data through the On-Premise Environments

In the case that the data asset provider has already installed an On-Premise Runner and he/she needs to upload data through it, the On-Premise Execution option should be selected as shown in Figure 4. After selecting the pre-processing steps that are to be applied on the uploaded data, by selecting Save, the Setup Harvester page for the On-Premise Execution appears, as shown in Figure 45. It needs to be noted that the only data loading option that the data asset provider has using the On-Premise Execution method, is to upload files at the moment.

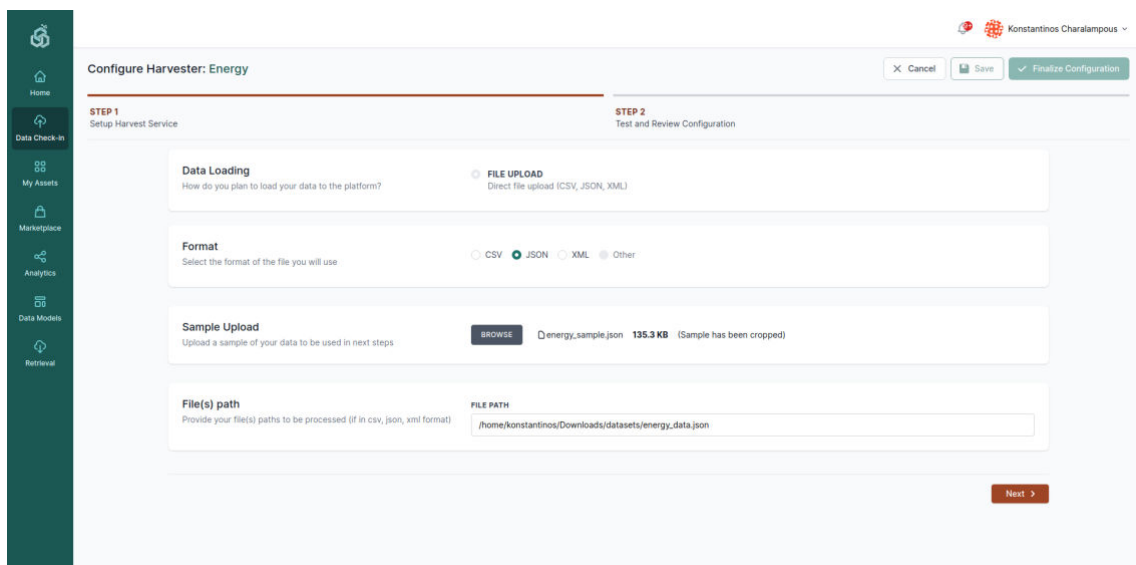


Figure 45: File Upload Method - On-Premise Execution - Setup Harvest Service (Step 1a)

Although the same procedure as in the File Upload Method in the Cloud Execution, for this case the data asset provider should include the full path to the file including the data, and in the next step the sample is uploaded in Step 1 and viewed in Step 2, as shown in Figure 46.

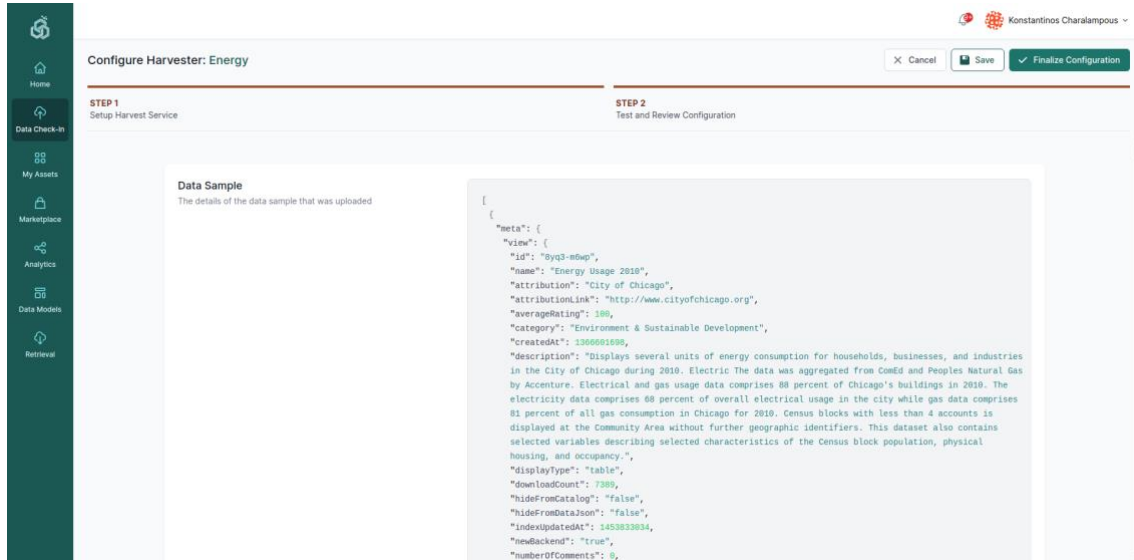


Figure 46: File Upload Method - On-Premise Execution - Setup Harvest Service (Step 2)

The rest of the steps are configured as presented in Section 3.2.

3.5 Manage Data Check-in Jobs

3.5.1 Edit a Data Check-in Job

Data asset providers are able to edit an existing data check-in job by navigating to the Data Check-in Jobs view, as shown in Figure 3. In particular, a data check-in job can be edited by selecting Edit from the options menu that is located at the right side of each data check-in job. Directly the Update Data Check-in Job view will be appear as depicted in Figure 47. Although the data asset provider is able to update the name and description of the data check-in job, the pre-processing rules and the data check-in execution location cannot be updated. By selecting the Update button at the top right part of this view, the data check-in job is updated and the data asset provider returns back to the Data Check-in Jobs view.

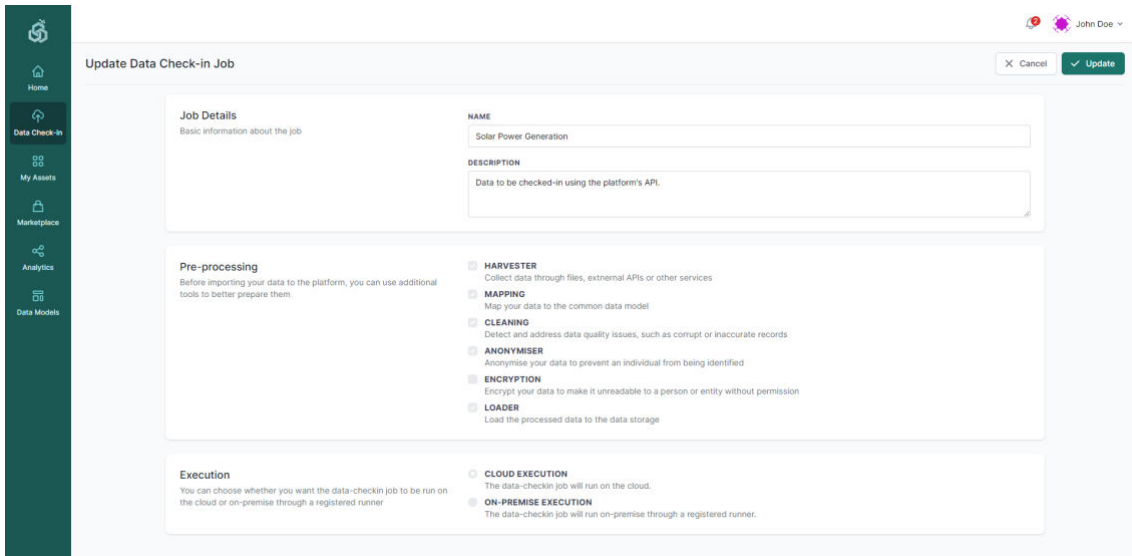


Figure 47: Update a Data Check-in Job

3.5.2 Append data to a Batch File Check-in Job

In the case of file upload as the selected harvesting method, the data asset provider may select “Append data” from the options menu at the right side of each data check-in job in the Data Check-in Jobs view. Such an action allows the user to upload additional file(s) whose data will be appended to the already checked in data as depicted in the following figure.

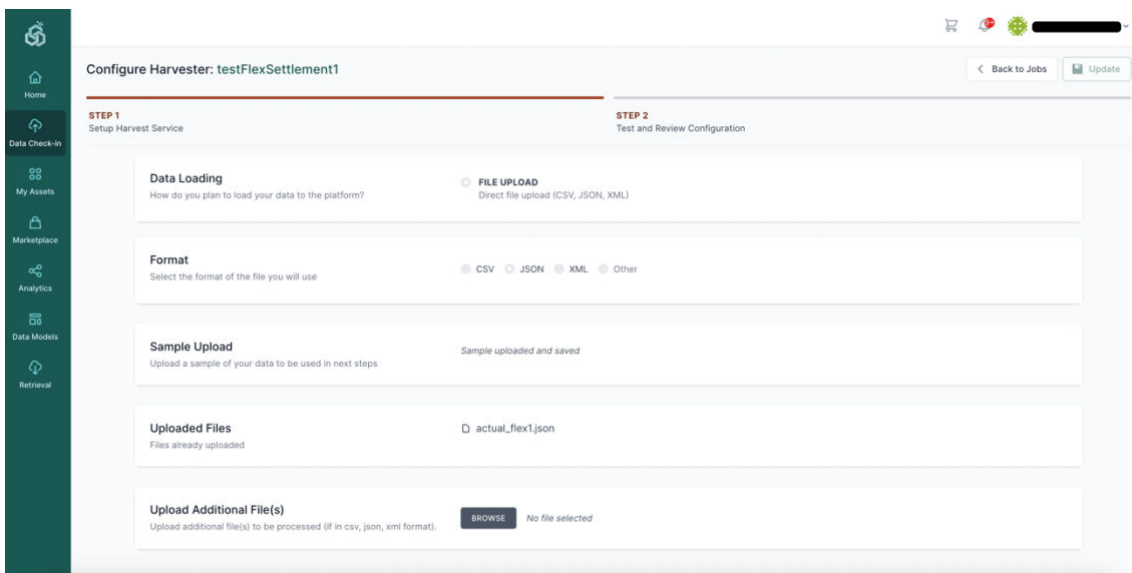


Figure 48: Append Data to a Batch File Data Check-in Job

3.5.3 Delete a Data Check-in Job

In the Data Check-in Jobs view, depicted in Figure 3 by selecting Delete from the options menu at the right side of each data check-in job, the data asset provider can delete a data check-in job. This will delete completely and permanently the configuration which means that if a data asset provider wants to execute again a data check-in job, he/she will not be able to do so. It needs to be noted that deleting a data check-in job does not affect the data that have been already stored in the SYNERGY Platform which can be deleted as described in Section 3.6.2.

3.5.4 View the Execution Logs of a Data Check-in Job

Data asset providers are able to view the overall execution history of a data check-in job by navigating to the Execution Log in the Data Check-in Jobs view. In particular, they can view the total number of executions, the number of successful and failed executions, the average execution time, and the details per execution as depicted in Figure 49.

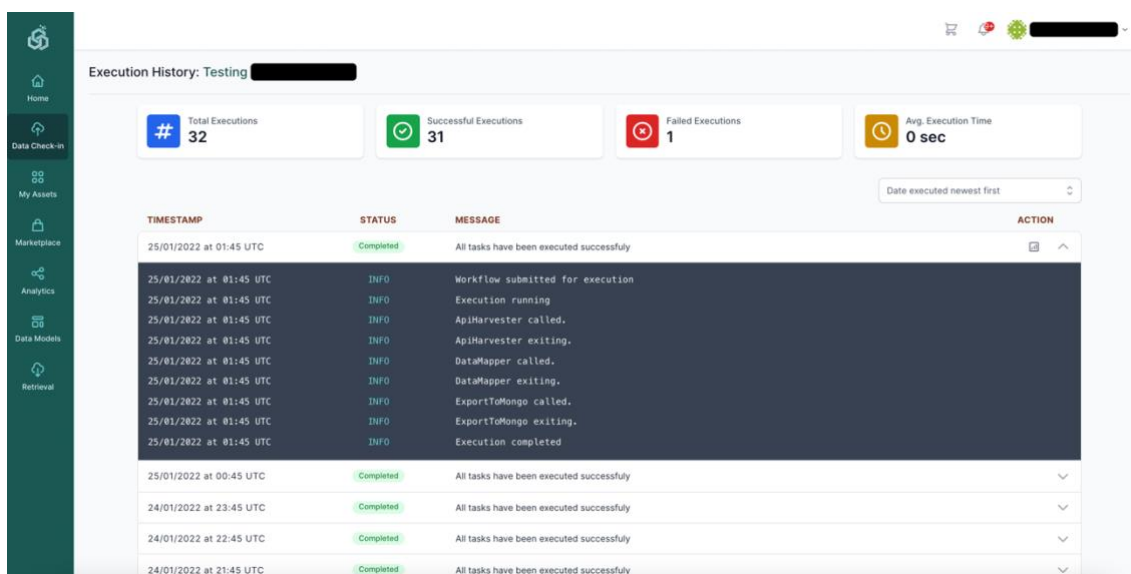


Figure 49: View the detailed execution history for a Data Check-in Job

3.6 Manage Data Asset Profiles

3.6.1 View all Data Asset Profiles

The data asset provider may view the list of all the data asset profiles that belong to his/her organisation or have been acquired by his/her organisation in the Assets View as shown in Figure 50, regardless its status (e.g. deprecated, incomplete, available). By selecting the options icon

located at the right side of each row, the data asset provider may delete, or edit a particular data asset profile. Editing a data asset profile, involves the same procedure followed during the definition of a new data asset profile as described in Section 3.2.3.

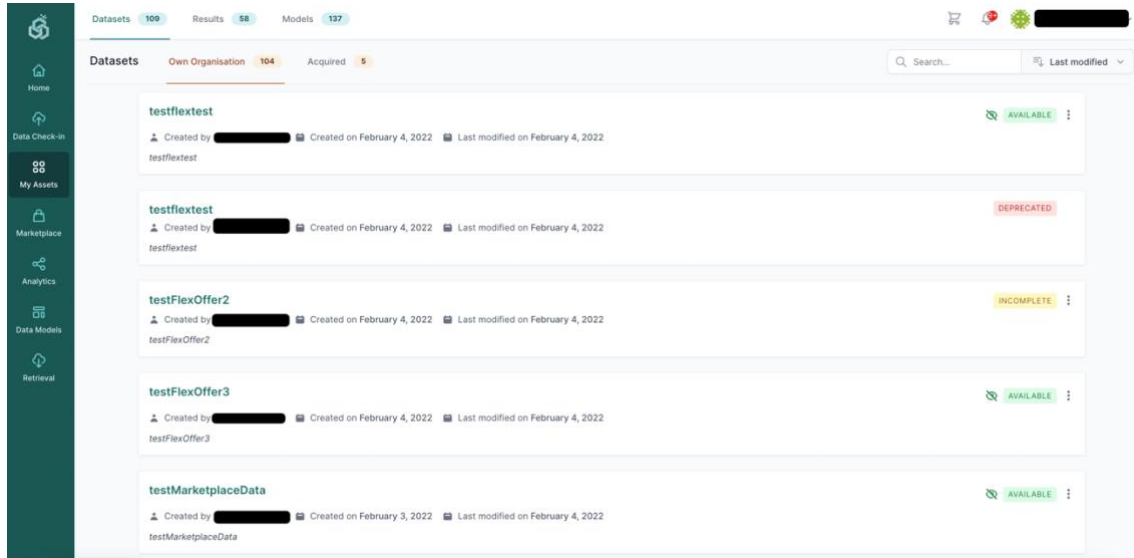


Figure 50: View all Data Asset Profiles

3.6.2 Delete a Data Asset Profile

A data asset provider may delete a Data Asset Profile by selecting Delete from the options menu located at the right side of each Data Asset Profile, as depicted in Figure 50. It needs to be noted that the actual data associated with the Data Asset Profile are deleted in case there is no active data asset contract. Any data retrieval operations associated with the specific dataset (as described in Section 4.4) are not available after the deletion of a particular Data Asset Profile, while the corresponding entry (in terms of metadata) will remain in the SYNERGY Platform as a deprecated Data Asset.

4 Data Search and Acquisition User Journey

A fundamental goal of the SYNERGY Platform is to provide all the functionalities needed for sharing and trading data assets in a secure and trustful manner, to stakeholders of the electricity data value chain. It needs to be noted that data assets can range from raw datasets, to models and analytics results.

Data asset consumers that are interested in acquiring data assets from data asset providers, reach an agreement stating the sharing terms, that is prepared based on the acquisition contract preparation process followed in the SYNERGY Platform. The acquisition contract preparation process is described in the subsequent subsections, while the different steps are depicted in Figure 51. In particular, data asset consumers may request for quotation for a particular data asset (or multiple data assets they have placed on their cart); the respective data asset providers receive the request for this data asset, check its content, and prepare a draft acquisition contract which is then sent back to the data asset consumer to accept, negotiate, or reject, accordingly. In the case of negotiation, this process is repeated until the data asset consumer and the involved data asset provider(s) reach consensus. Then, the data asset consumers need to settle the acquisition contract through a crypto-currency payment that is supported in the SYNERGY Platform (and leads to the immediate remuneration of all involved data asset providers in the case of datasets, models or results, and activation of the contract), or an offline payment (that needs to be manually confirmed by the respective data provider in order to activate the contract in the SYNERGY Platform, in the case of datasets only).

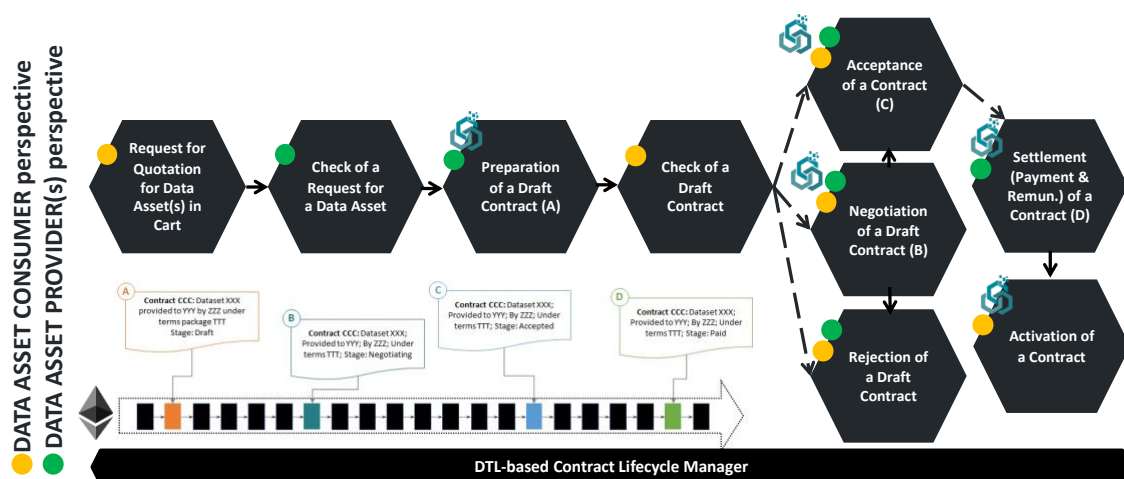


Figure 51: Contract Preparation Process

Once the data asset consumer has an active acquisition contract with the provider, the acquired data asset is available for use in the SYNERGY Platform (in the consumer's organization Secure Experimentation Playground). It needs to be noted that both stakeholders of the electricity data value chain and the SYNERGY energy apps developers are expected to utilize the data search and acquisition user journey to get access to data assets from different stakeholders in the SYNERGY Platform.

If multiple data assets are acquired at once through the cart functionality, the separate contracts that are negotiated per asset cannot be activated unless all involved parties (for each data asset) have reached an agreement. Essentially, this means that a rejection for a request of contract of a data asset (irrespective of the extent to which the process for all the rest of the data assets has proceeded) leads to the cancellation of the process for all assets within the cart.

In the case of models and analytics results, it is expected that they may have emerged as derivative data assets from different data asset providers. In order to ensure that any derivative asset that is to be available in the SYNERGY Marketplace, has the explicit approval of the original data asset providers, a different type of contract, namely a derivation contract, needs to be signed between the involved parties (on top of any existing acquisition contract) to clarify the expected reimbursement of the original data asset providers when their assets are thus indirectly shared/sold.

In this section, the core three components involved in the Data Sharing Services Bundle, namely (a) the Data & AI Marketplace, (b) the Contracts Lifecycle Manager, and (c) the Remuneration Engine are described through the Data Search and Acquisition User Journey.

4.1 Navigate to the SYNERGY Marketplace

Data asset consumers may access the SYNERGY Data & AI Marketplace where they can navigate through the data assets that they are eligible to view (upon satisfying the access policies set by their respective providers).

As depicted in Figure 52, data asset consumers view highlight info regarding each data asset such as the type of data asset (i.e., dataset, model, result), the title of the data asset, a short description regarding the data asset, the organisation that owns or manages the data asset, and a cover image related to the data asset. It needs to be noted that data assets that are stored locally are considered as not to be shared and are not shown in the SYNERGY Marketplace.



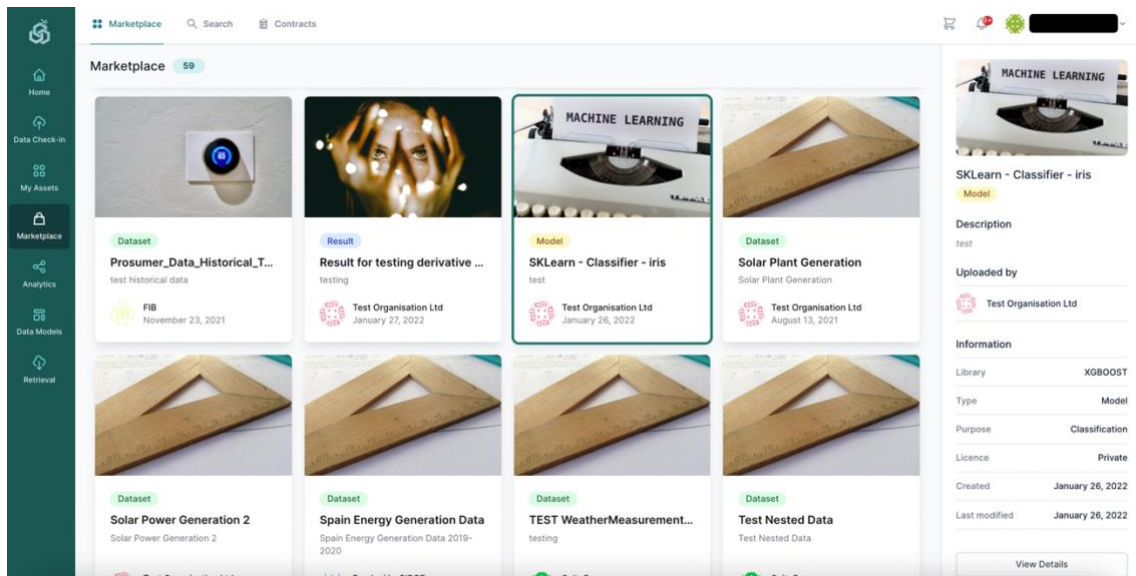


Figure 52: SYNERGY Data & AI Marketplace

Data asset consumers navigating through the SYNERGY Marketplace, can potentially search for a particular data asset by typing in the search bar, and/or by browsing data assets by different parameters (e.g. type, and depending on the type selection: domain, category, accessibility, format, and language). The corresponding results from the search functionality are filtered and displayed directly in the Results View, as depicted in Figure 53, where one can sort the results based on relevance, title, and date available/modified.

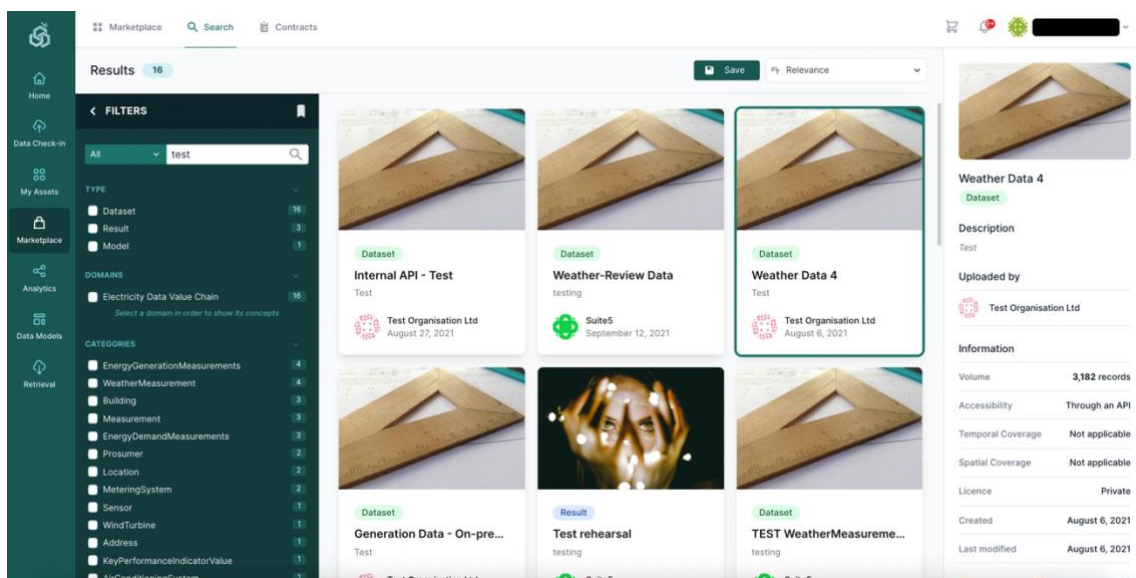


Figure 53: Search Page

If the data asset consumer selects a particular data asset in the Marketplace, he/she will be redirected in a new browser tab where the overview and more detailed information of the data asset depending on its type (as described in Section 3.1.3 for datasets, and in Section 5.5 for models and results) will appear.

Additionally, a data asset consumer may perform advanced, targeted data queries by selecting a domain of his/her interest which reveals the “Show concepts” button that enables the Concepts slideover as shown in Figure 54. Hence the data asset consumer may select multiple fields/concepts to build a data query, which enables the Data Query slideover for adding, editing, or removing conditions to the query.

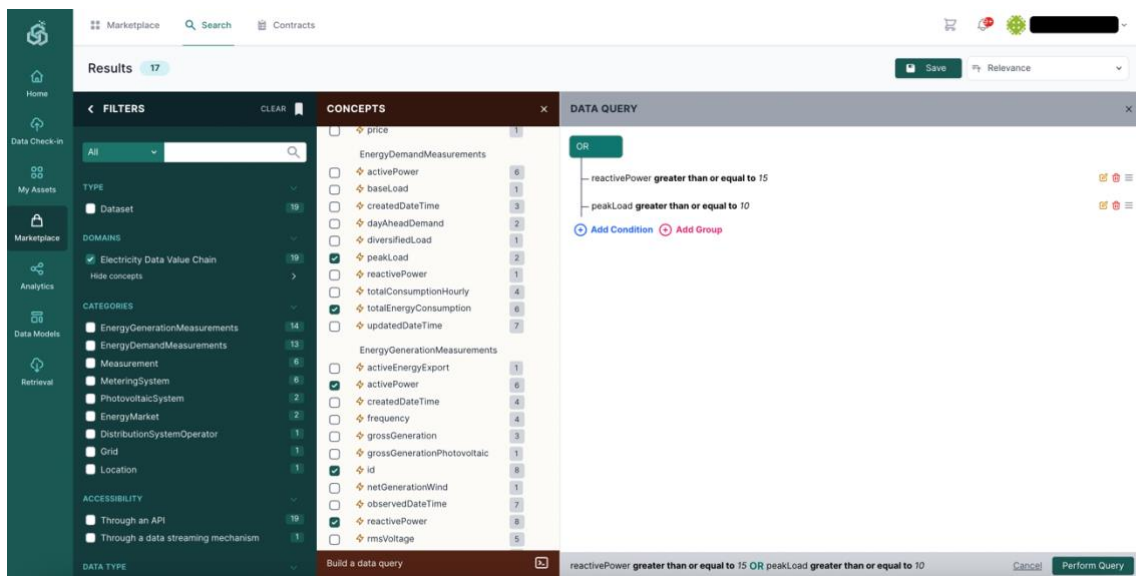


Figure 54: Search Page - Advanced Data Query

Selecting to “Perform a Query”, a search within the selected concepts and the added conditions is triggered, and the corresponding results are shown.

4.2 Acquire a single Data Asset

As organisations in the SYNERGY Platform may undertake the role of both data asset providers and consumers, their members may view the contracts of the data assets they are involved in (as data asset providers or consumers), search for a particular contract, filter the contracts based on their status, and sort them, as shown in Figure 55. Such contracts are classified into: (a) acquisition contracts (for determining the sharing terms of a private data asset), (b) derivation contracts (for reaching consensus prior to derivative data assets being made available in the

SYNERGY Marketplace), and (c) public asset contracts (in order to have visibility who has asked for access to data assets that are provided as Public). In general, the contracts are visible in a list view, while for each contract, basic information such as the data asset title, the contract details (i.e., date created, date last updated, data asset provider, data consumer), the contract status, and any available action.

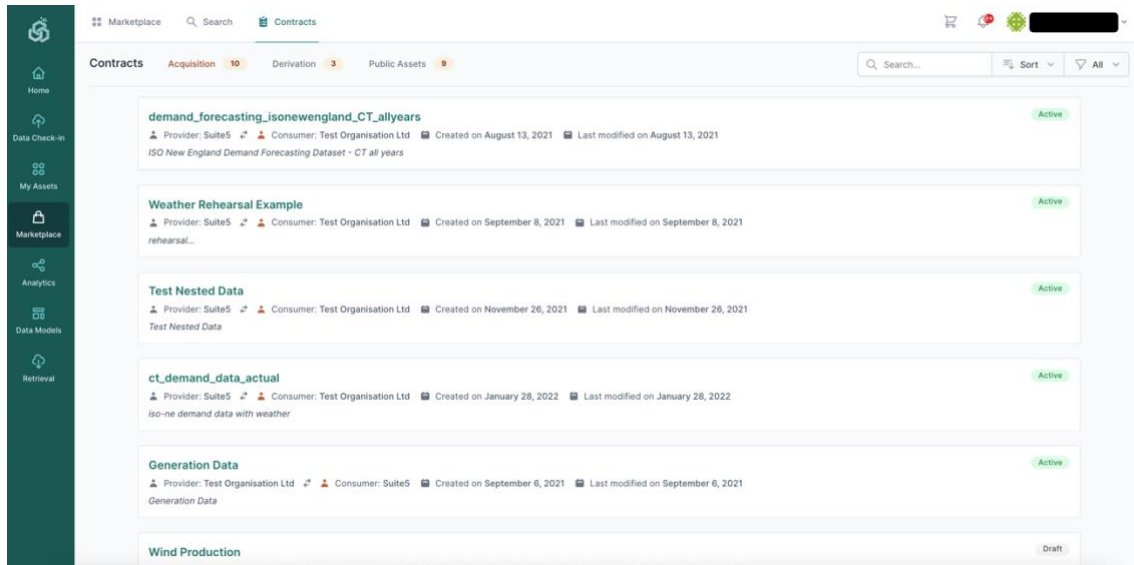


Figure 55: SYNERGY Data & AI Marketplace – Contracts List

4.2.1 Request a single Data Asset (Data Asset Consumer Perspective)

Data asset consumers may initiate the process for formally acquiring a data asset by creating a request to the data asset provider as shown in Figure 56. In particular, the data asset consumer needs to insert the details regarding the data asset request, and specifically to select the duration of use of the data asset (including any updates available during the respective period), to select whether the data asset will be used exclusively in the SYNERGY platform or if the data consumer would like to download the data asset “locally” through the SYNERGY Platform APIs. In addition to these, the data asset consumer should optionally insert a message to be provided to the data asset provider, describing how the data asset will be used. Then, in case the data asset is a dataset, the data asset consumer may view the data asset structure and the processing rules that have been applied on the data, and select the exact fields to be acquired, if not the whole data asset. During this step, the data asset consumer may also apply filters to the (unencrypted) fields to obtain a data asset “slice”, whenever applicable. Once the data asset

consumer is satisfied with the details inserted regarding the request of the data asset, he/she is able to submit the request to the organisation’s manager at the data asset provider side.

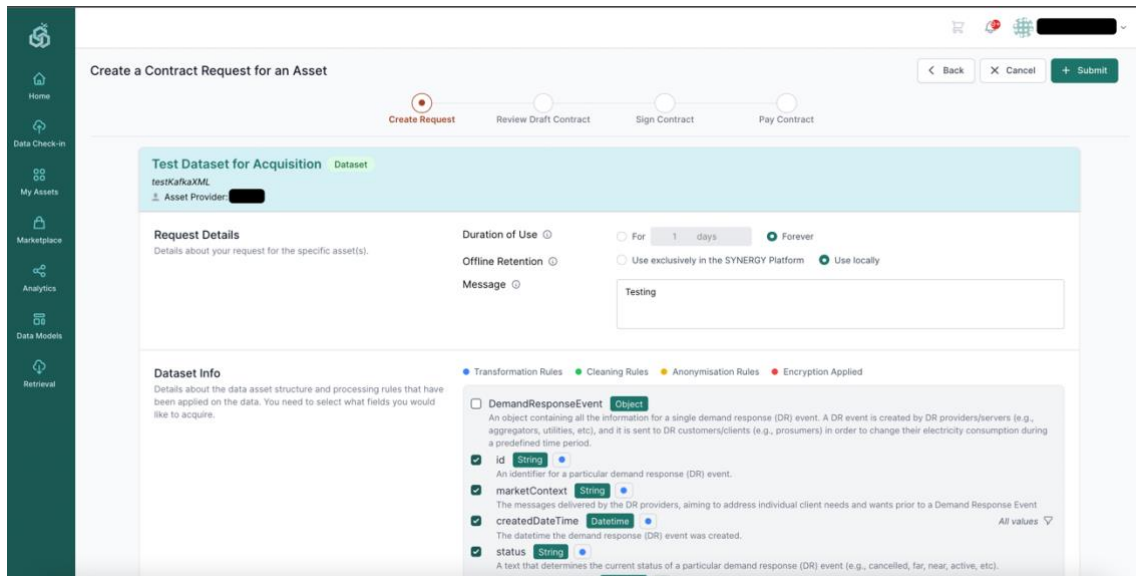


Figure 56: Request for Data Asset Acquisition by the Organisation's Manager (Consumer Perspective)

4.2.2 Review a Data Asset Request (Data Asset Provider Perspective)

The manager of the organisation (acting as the data asset provider) receives a notification (in the platform and/or via email) and is able to review the details of each request made by a potential data asset consumer. He/she proceeds to reject or accept the request for the data asset based on the information provided, as depicted in Figure 57.

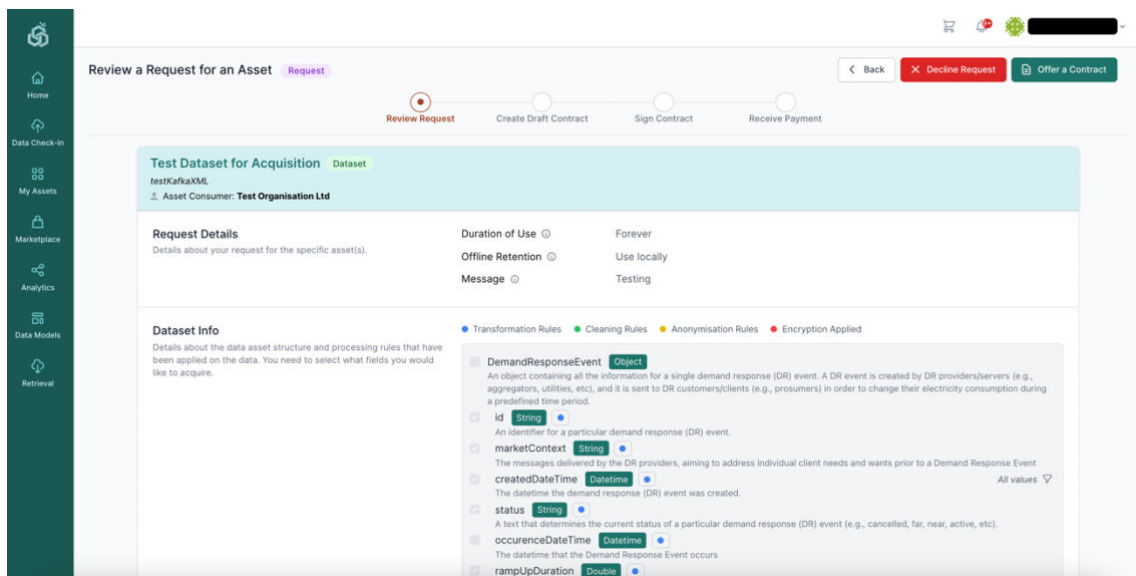


Figure 57: Review Request for Data Asset Acquisition (Provider Perspective)

4.2.3 Prepare a draft contract (Data Asset Provider Perspective)

In case the organisation’s manager acting as data asset provider has accepted the request, a draft contract for the corresponding data asset is prepared as shown in Figure 58. The manager may review once more the details of the data asset (depending on its type), define the payment details, and prepare the contract terms that concern the data asset acquisition between the data asset provider and consumer. It needs to be noted that the terms that are enforced by the SYNERGY Platform are also listed and cannot be edited by any party.

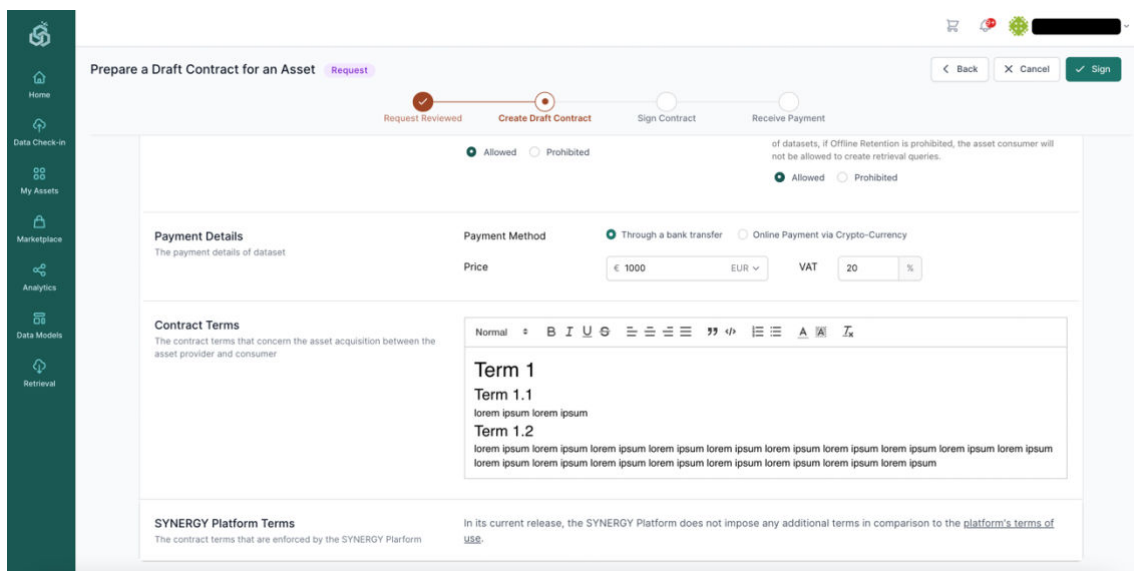


Figure 58: Create Draft Contract for Data Asset Acquisition by the Organisation’s Manager (Provider Perspective)

The organisation’s manager is then able to sign the draft contract upon inserting the password of the organisation’s wallet (which he/she manages), in order to unlock the wallet and sign the draft contract in the blockchain.

4.2.4 Review a draft contract (Data Asset Consumer Perspective)

Once a draft contract is signed by the data asset provider organisation, the manager of the data consumer organisation is notified (in the SYNERGY Platform and/or via email) and may view the contract details, as shown in Figure 59. Upon reviewing the details of the draft contract prepared by the data asset provider, the manager of the data asset consumer may accept, negotiate or reject the offer accordingly. During this step, the data asset consumer is also able to download the draft contract as a pdf file.

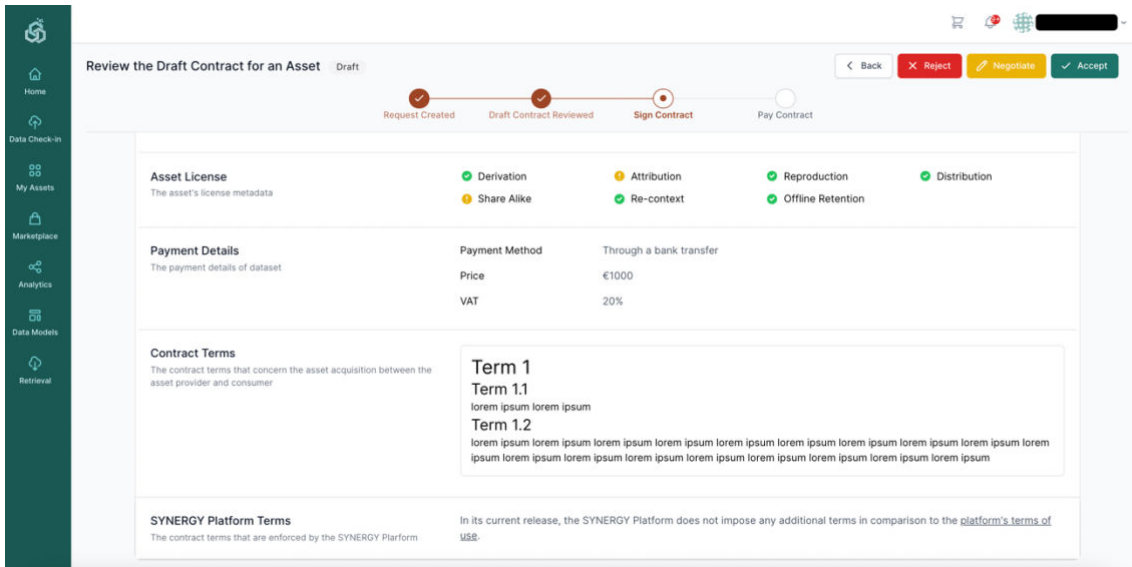


Figure 59: Review Draft Contract for Data Asset Acquisition (Consumer Perspective)

If the manager representing the data asset consumer has opted to accept the draft contract, he/she is prompted to insert the organisation’s wallet password and write the contract to the SYNERGY blockchain (as depicted in Figure 60). The process continues as described in Section 4.2.6.

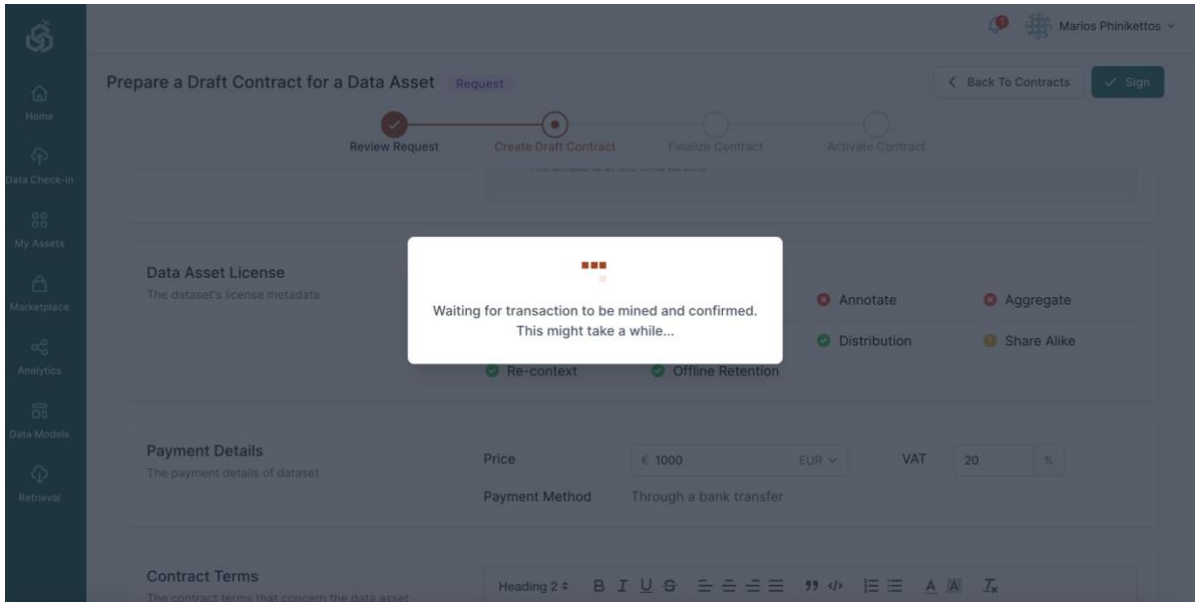


Figure 60: Sign a Contract for Data Asset Acquisition (Consumer/Provider Perspective)

On the contrary, if the manager representing the data asset consumer has rejected the draft contract’s terms, then the process concludes, and the data asset consumer does not get access

to the requested data asset. The data asset provider also views the contract with “rejected” status.

4.2.5 Negotiate a draft contract (Data Asset Consumer Perspective)

In case that the data asset consumer chooses to negotiate any part of the contract, e.g. regarding the license, the payment and/or the terms (as displayed in Figure 61), he/she is able to change the payment details such as the cost, as well as update the contract terms by adding new terms, editing or removing the existing ones. Following that, the data consumer manager will be able to sign the revised contract, by inserting the organisation’s wallet password and writing the contract to the SYNERGY blockchain.

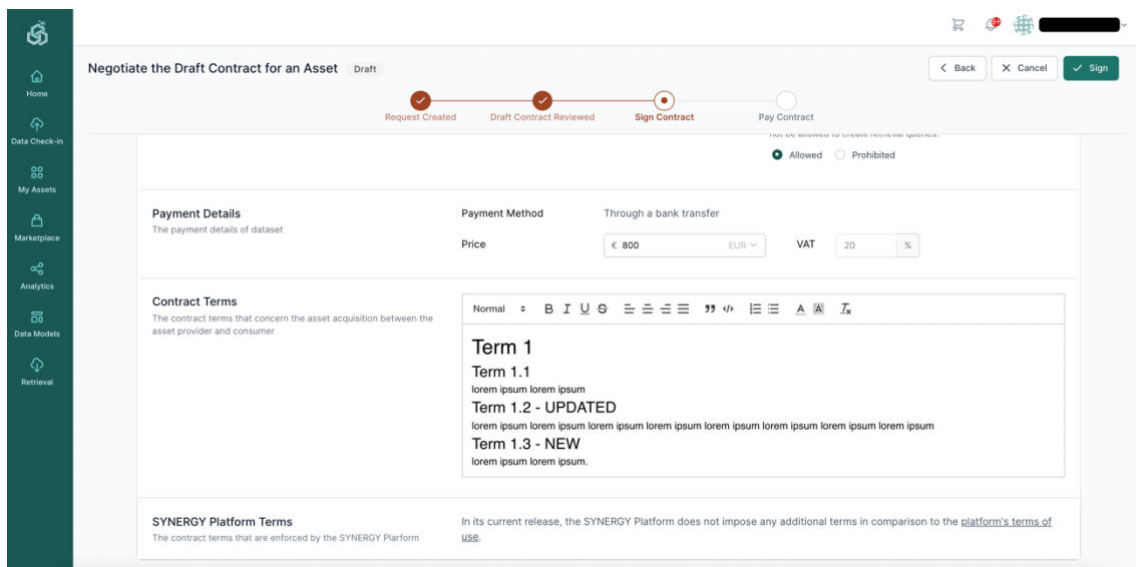


Figure 61: Negotiate Draft Contract for Data Asset Acquisition (Consumer Perspective)

4.2.6 Review a revised contract (Data Asset Provider Perspective)

Once a contract is revised, the data asset provider manager is notified to review the changes that the data consumer asked for, as shown in Figure 62. As it happened in the case of the draft contract on behalf of the data asset consumer, though, in Section 3.3.4, a data asset provider is able to accept, reject or further negotiate the revised contract.

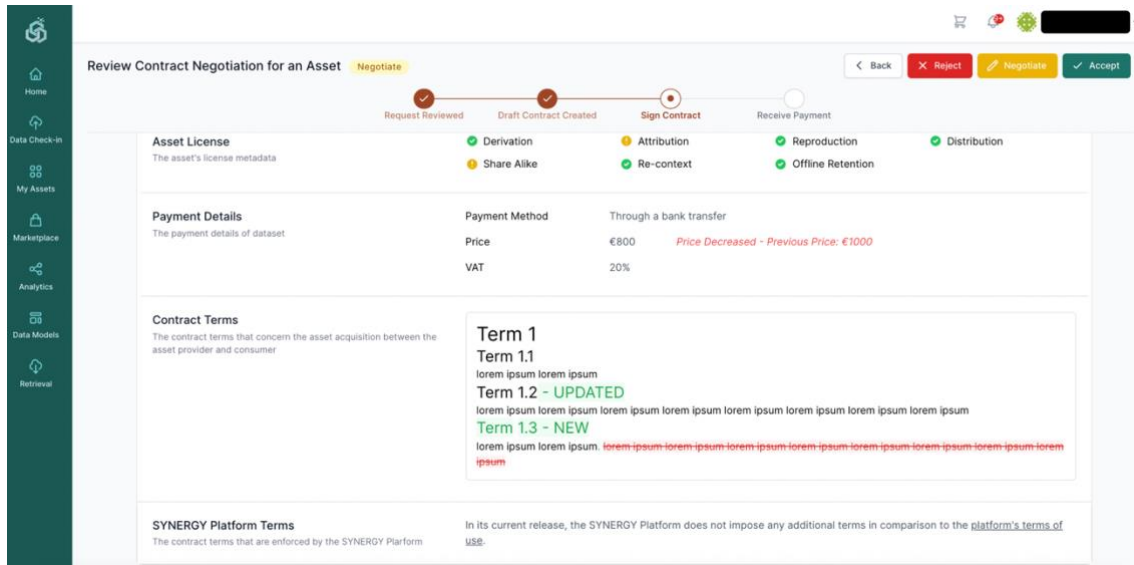


Figure 62: Review Updated Draft Contract for Data Asset Acquisition (Provider Perspective)

It needs to be noted that this procedure described in Sections 4.2.4, and 4.2.5 may iterate until both the data asset provider and data asset consumer are satisfied with the contract for the data asset acquisition.

4.2.7 Settle a finalized contract (Data Asset Provider Perspective)

Once the data asset provider and the data asset consumer have reached consensus and signed the same version of the contract, the respective payment, if any, needs to be settled according to the payment method that is mentioned in the signed contract. In the case of single datasets, the data consumer is allowed to pay offline (i.e. not in the crypto-currency supported in the SYNERGY Platform) the price of the data asset as dictated in the contract’s terms, and the data asset provider needs to manually confirm the payment in order for the contract to be activated as shown in Figure 63.

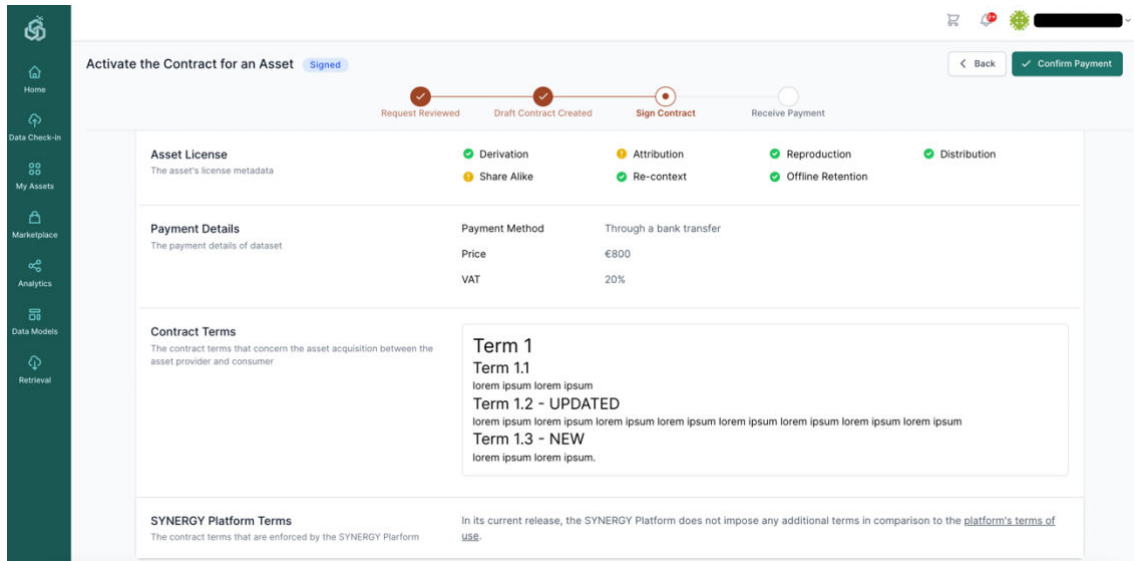


Figure 63: Activate Draft Contract for Data Asset Acquisition (Provider Perspective)

As soon as the contract has been activated by the data asset provider, the data consumer is eligible to acquire the data asset according to the contract's terms, e.g. the SYNERGY platform transfers the data asset to the data asset consumer's Secure Experimentation Playground (for use only in the SYNERGY platform).

4.3 Acquire multiple Data Assets through the Cart

4.3.1 Request multiple Data Assets (Data Asset Consumer Perspective)

Data asset consumers are able to add multiple data assets in their cart (as depicted in Figure 64) and decide how they prefer to acquire them: (a) each data asset separately (as described in section 4.2), or (b) all data assets (up to five) as a bundle (which means they acquire them all or none).

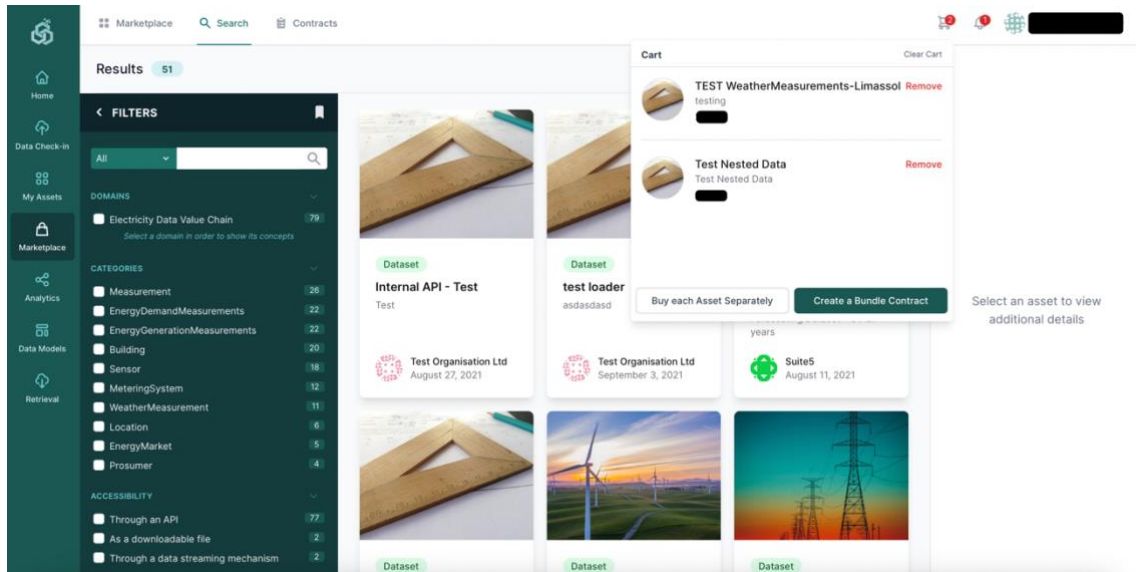


Figure 64: Add data assets of interest in Cart (Consumer Perspective)

Once the data asset consumers decide to create a bundle contract, they are able to create the respective requests to the data asset providers, declaring the duration of use, whether offline retention is allowed and a message describing the request as depicted in Figure 65.

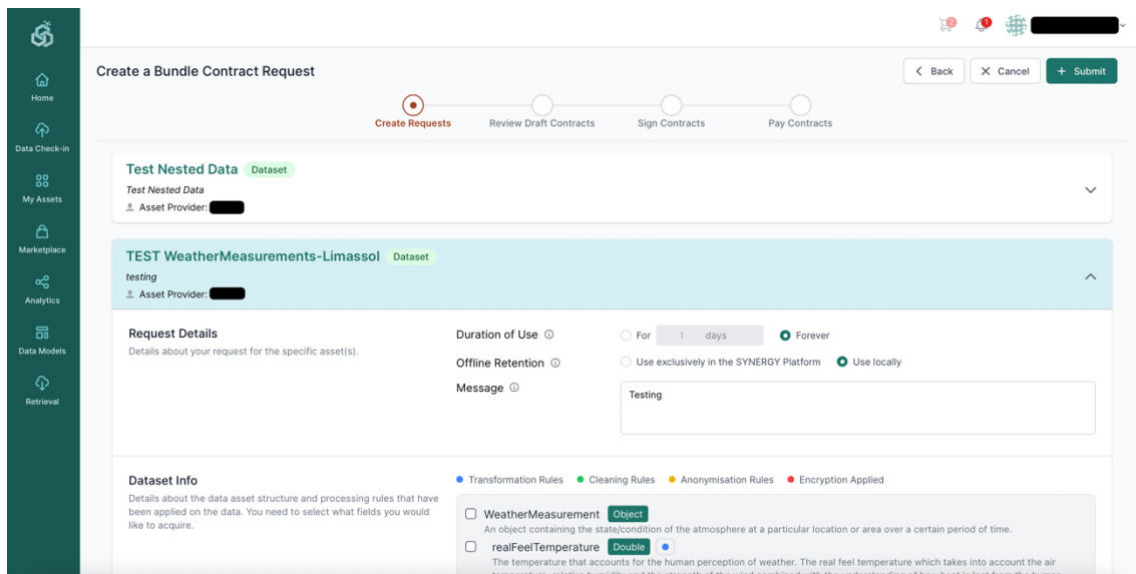


Figure 65: Create a request for a bundle of data assets (Consumer Perspective)

In contrast to the single asset requests, a bundle request needs to be directly written to the SYNERGY Blockchain, thus the data asset consumer (i.e. manager of the organisation) is prompted to provide the wallet’s password and the transaction is mined as depicted in the following figure.

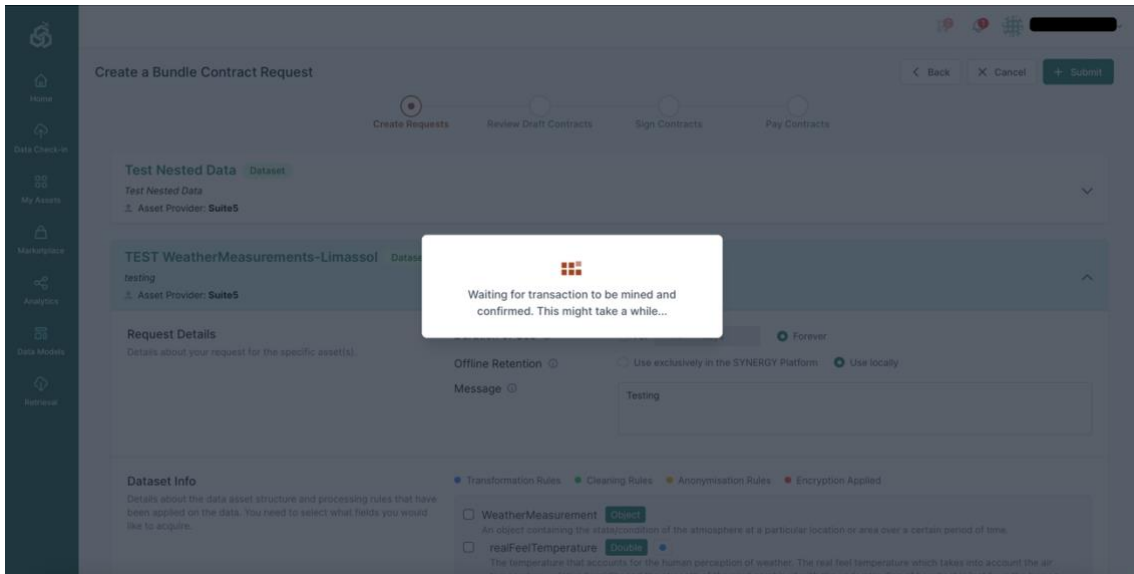


Figure 66: Writing the bundle request in the SYNERGY Blockchain (Consumer Perspective)

4.3.2 Managing the Contracts in a Bundle (Provider & Consumer Perspectives)

The managers of the organisations (acting as the data asset providers) receive appropriate notifications (in the platform and/or via email) and are able to review the details of each request made by a potential data asset consumer. They are able to proceed to reject or accept the request for the data asset based on the information provided, as depicted in the following figure. It needs to be noted that in this example, the bundle request referred to 2 data assets from the same data asset provider.

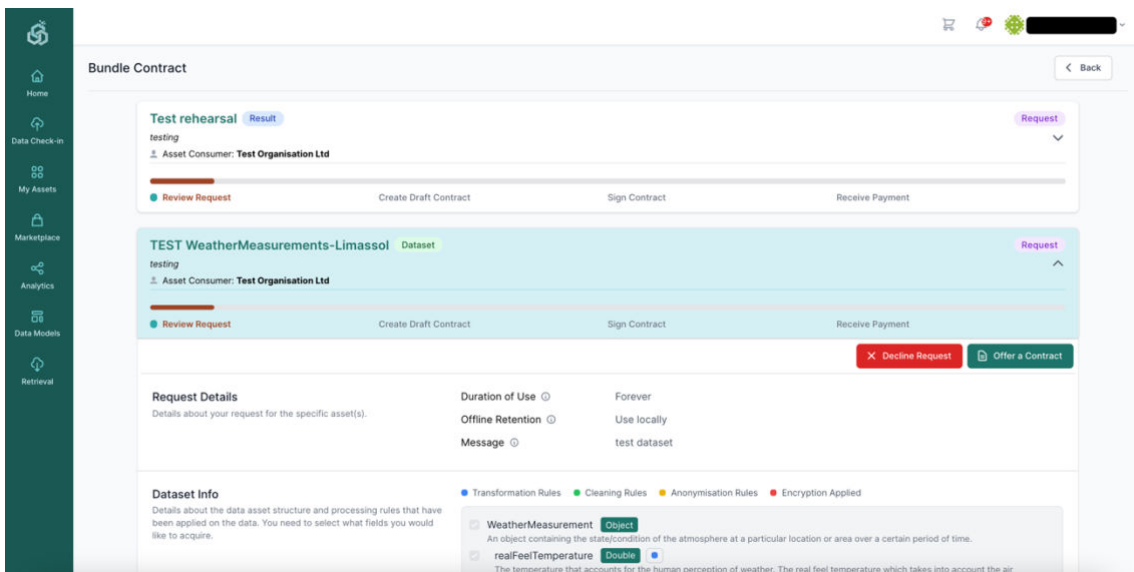


Figure 67: View the Bundle Contract Requests (Provider Perspective)



From this point onwards, the data asset provider of each data asset decides whether they want to offer a contract for the specific bundle’s asset and the process proceeds as described in sections 4.2.3-4.2.6 for each asset. If any of the data asset providers declines a contract, the bundle cannot be acquired and all contracts for the assets of a bundle are cancelled.

At any moment, the data asset consumer has visibility about the status of the contracts for the assets belonging in the bundle as depicted in the following figure.

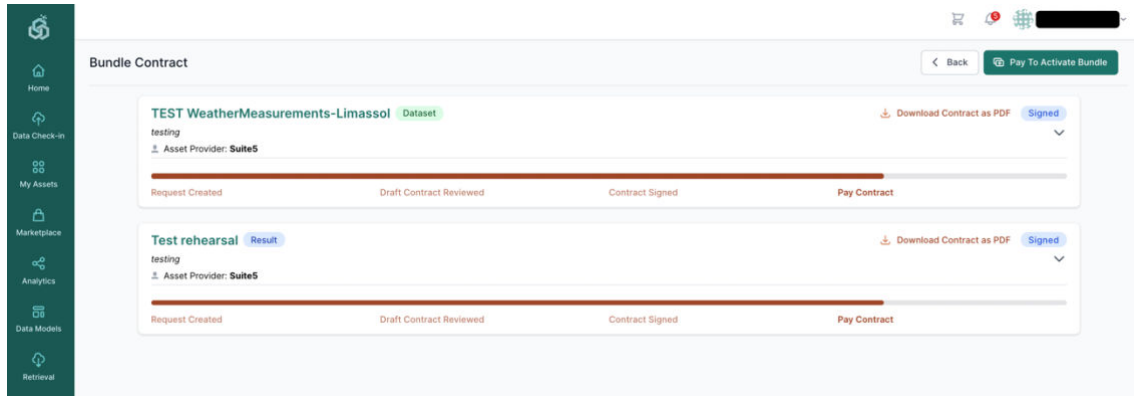


Figure 68: Review the status of the bundle acquisition (Consumer Perspective)

4.3.3 Settle the Contracts in a Bundle through Crypto-Currency (Consumer Perspectives)

Once the contracts for all data assets in the bundle are signed, the data asset consumers need to pay the total price through the crypto-currency supported in SYNERGY (ETH). They need to unlock the wallet and pay the respective price as depicted in Figure 69. The funds are immediately transferred to each data asset provider’s wallet and the respective contracts are automatically activated (without requiring any further action by any involved party).

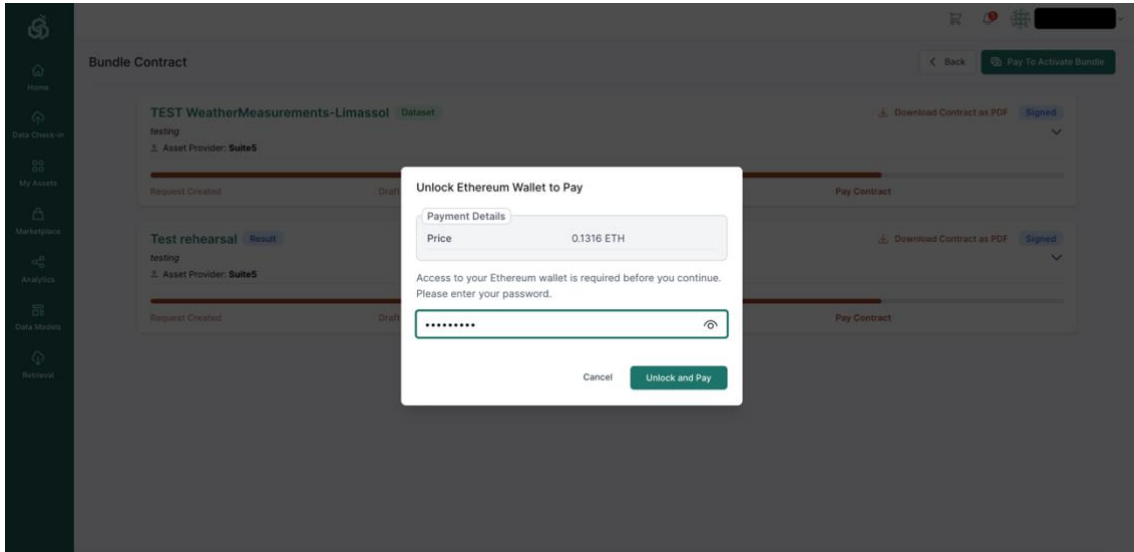


Figure 69: Payment through Crypto-Currency (Consumer Perspective)

4.4 Retrieve a Data Asset

4.4.1 View all Saved Retrieval Queries

A data asset consumer may view all the retrieval queries that have been saved in the SYNERGY Platform by its organisation’s members, by accessing the Retrieval menu. The data asset consumer is allowed to rename or delete a saved query as well as access the saved query configuration and the data assets that it includes. In the latter case, he/she can view again the current configuration or proceed with updating it.

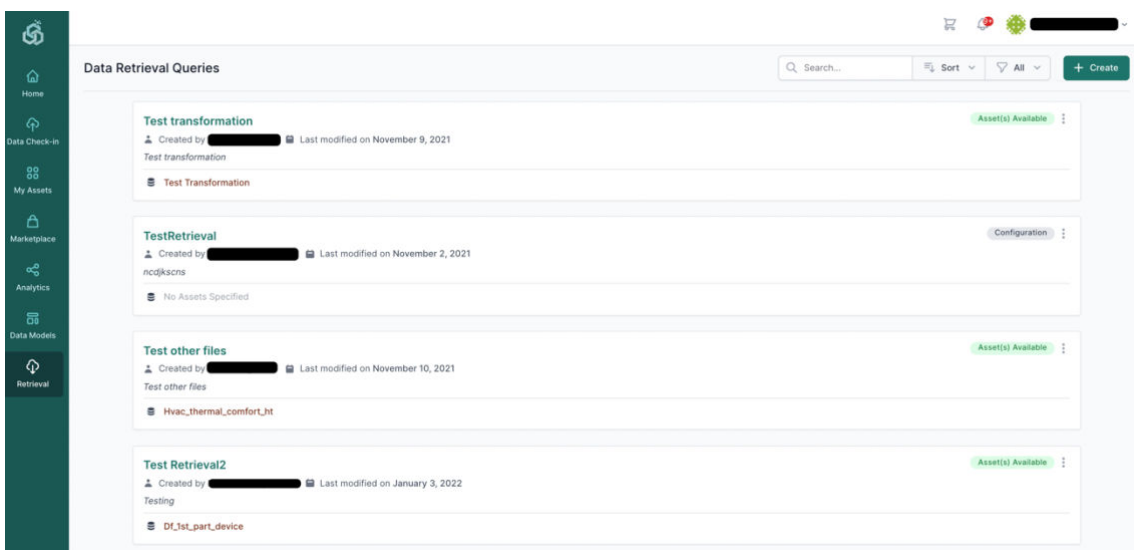


Figure 70: View Saved Retrieval Queries

4.4.2 Create a Retrieval Query

Data asset consumers may create a new Retrieval Query defining its title, description and acquisition method for datasets and results (that belong to their organisation or which their organisation has acquired), as described in the following figure.

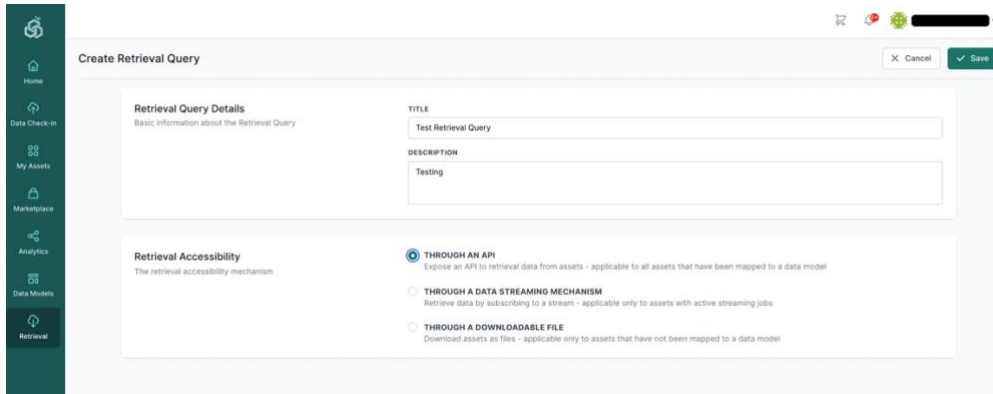


Figure 71: Creation of a retrieval query

As depicted in Figure 72, they need to select the concepts from the selected data assets that will be extracted and are expected to be returned as the retrieval query results. In addition, the concepts that will be used as query parameters to filter the query results need to be defined (with different options becoming available depending on the concept’s data type).

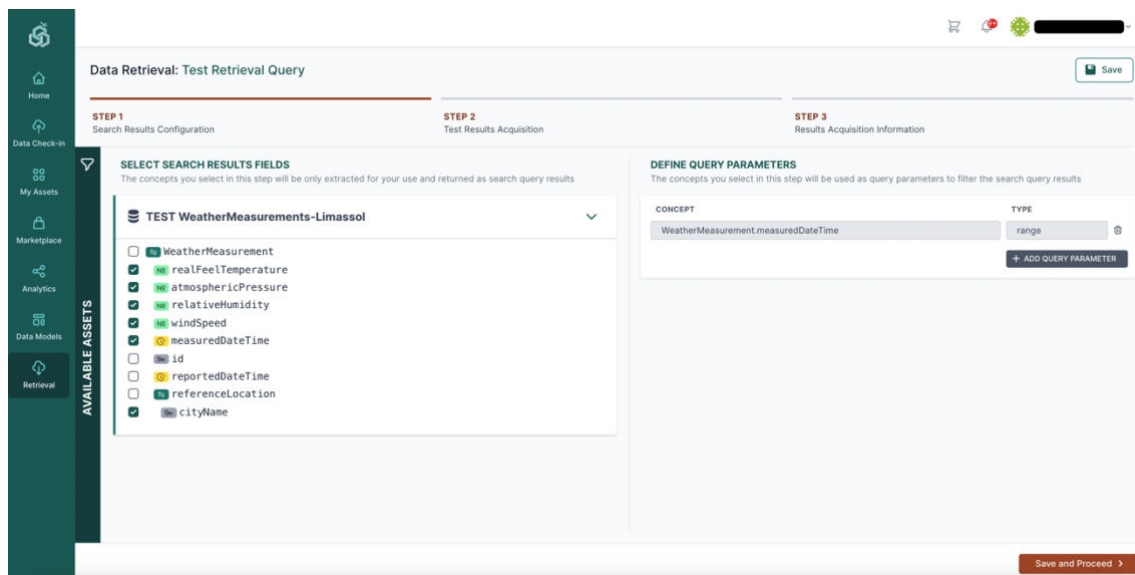


Figure 72: Retrieval Query Configuration (Step 1)

Following that, as depicted in Figure 73, at the left side of this page, the data asset consumers shall be able to edit the body with indicative values for the concepts/fields, to be used as the

query parameters in order to appropriately filter the dataset, while at the right side of the page, a sample preview of the result is shown after selecting “Run Query” to acquire the sample of the query results.

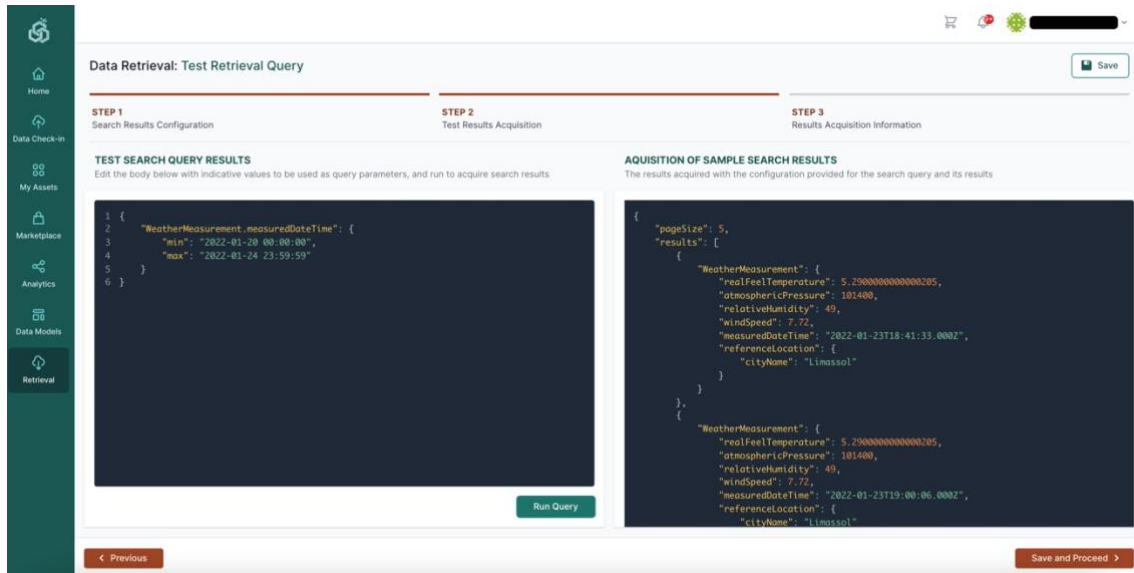


Figure 73: Retrieval Query Configuration (Step 2)

The last step of the retrieval configuration instructs the data asset consumers about how to acquire the retrieval results through the SYNERGY APIs as depicted in Figure 74. Instructions regarding authentication, pagination and sorting of results are provided in addition to the endpoints (i.e., for GET and POST methods) including the full API paths containing the base URL and the different path segments to be used for the query, and the query body (for the POST method). It needs to be noted that the platform requires the user requesting the retrieval of data from an API, and receiving the associated query results, to be authenticated by adding an already generated retrieve access token into an appropriate header of the request. In case that the data asset consumer does not have an already generated token, he/she needs to generate one by selecting the “generate a new one” link, and then to follow the same procedure to add it into the request header. In addition, different options for sorting and paginating the results are available. For sorting the results in a specific order, the data asset consumer needs to provide sorting query parameters (e.g. orderBy that specifies the full path of the field to be ordered by, and orderDirection that specifies the ordering direction - ASC for ascending ordering - and DESC for descending ordering) in the request. Finally, data asset consumers are able to retrieve the measurement units for the fields included in the retrieval query through a customised endpoint.

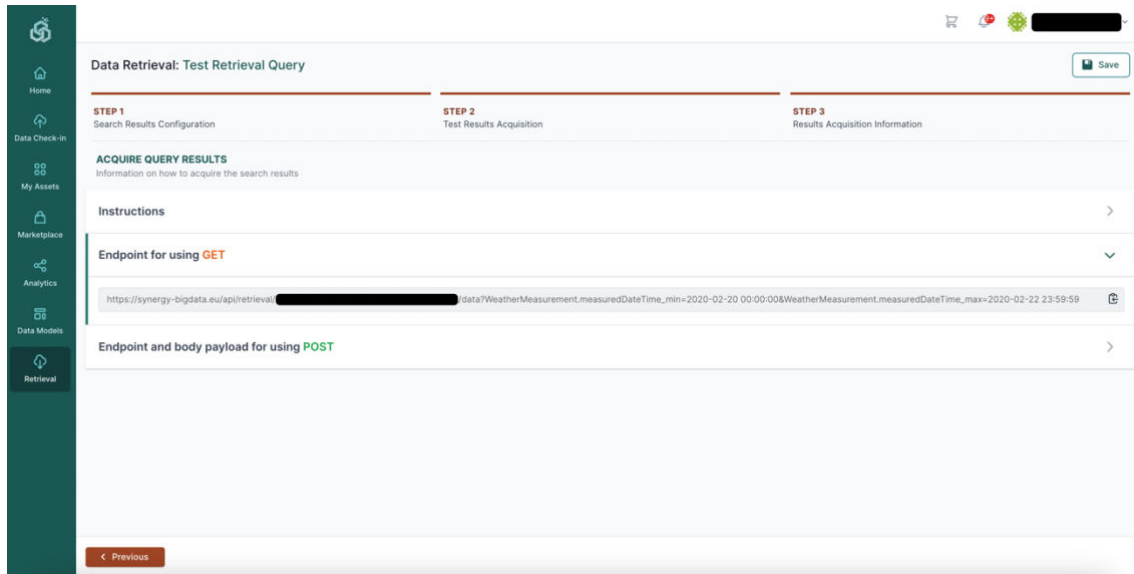


Figure 74: Retrieval Query Configuration (Step 3)

It needs to be noted that if any of the data assets appearing in a retrieval query has been deprecated or their respective provider(s) have revoked access to them (through changes in the access policies or the involved contract has expired), a relevant notification is provided to the data asset consumer (in the SYNERGY Platform and/or via email).

5 Data Analytics User Journey

The SYNERGY Platform provides the Data Analytics Services Bundle, described in D4.3 “SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles – Release 1.00”, that offers all the functionalities around data processing and analysis to gain valuable insights over data assets that have been checked-in in the SYNERGY Platform as discussed in Section 3, or have been acquired from other electricity data value chain stakeholders (according to the Data Search & Acquisition User Journey described in Section 4). As mentioned in D4.3 “SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles – Release 1.00”, there are five main components included in the Data Analytics Services Bundle, namely: (a) the Analytics Workbench, (b) the Visualization & Reporting Engine, (c) the Data Manipulation Service, (d) the Analytics Execution Service, and (e) the Secure Results Export Service. These components are described in this section divided into different workflows (i.e. Create, Configure and Execute an Analytics Pipeline, Register a pre-trained Model, Make a derivative asset available in the SYNERGY Marketplace, and Visualise the Results of an Analytics Pipeline). Similarly, as in the case of Data Check-in workflow, there is a distinct separation between the design of a data analytics pipeline and its execution either in the organisation’s Secure Experimentation Playground in the cloud, or on-premise. Although both data asset providers and data asset consumers may proceed to the Data Analytics User Journey, they will be referred to as data asset consumers, for brevity. It needs to be noted that both stakeholders of the electricity data value chain and the SYNERGY energy apps developers are expected to utilize the data analytics user journey to run analytics on the SYNERGY Platform.

5.1 View all Analytics Pipelines

The data asset consumers may see all the analytics pipelines they have defined as depicted in Figure 75, where the details, the status and the available actions for each analytics pipeline is displayed. A data asset consumer may sort the list of analytics pipelines by the title, the date it was created, the date it was executed, the latest execution(s) status (since a pipeline may be executed multiple times according to a schedule) and the user (within the organisation) who created it. Clear indications appear when the execution workflow definition of the pipeline is still incomplete or not valid, as well as in the cases of pending pipelines that are scheduled to be executed in the future, while successful and failed executions are highlighted with appropriate



icons. Finally, a data asset provider may select to configure or remove a pipeline, to view the results of the pipeline in the pipeline results view (as described in Section 5.6), and view the pipeline execution log, from the Actions column.

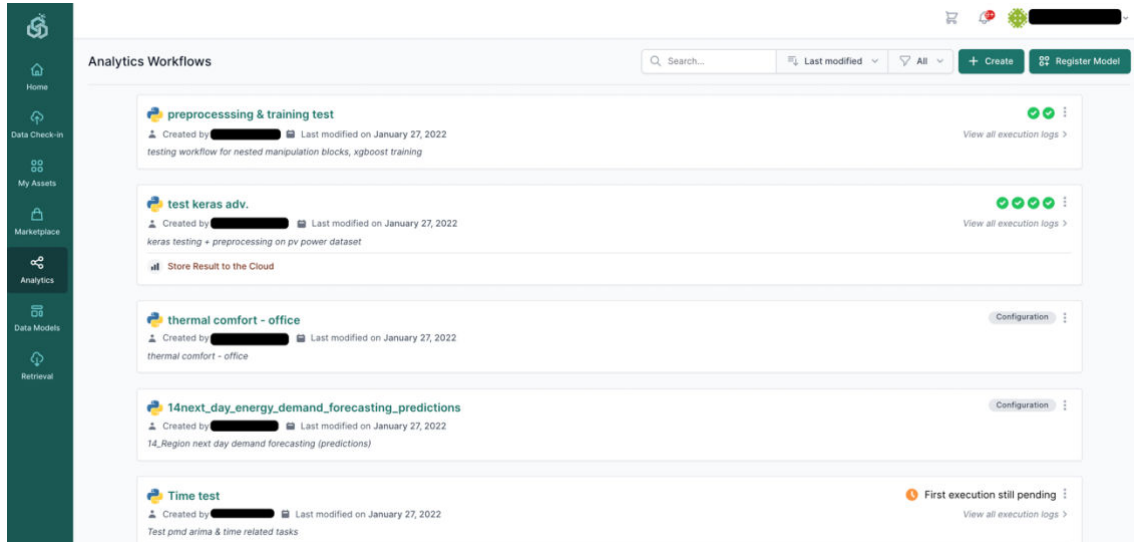


Figure 75: View all Analytics Pipelines

5.2 Create an Analytics Pipeline

A data asset consumer may create a new analytics pipeline through the Data Analytics view. Initially, the view for creating a new Workflow appears as depicted in Figure 76, where the data asset consumer needs to insert some basic details such as a title, and a short description that essentially provides an overview of this data analytics pipeline job. Then, the data asset consumer needs to select the execution framework of their analytics pipeline, that is the framework (i.e., Apache Spark engine, or Python environment) that will be used for the data processing tasks included in the pipeline. In addition, the location where the analytics pipeline job will be executed (i.e., Cloud Execution, or On-Premise Execution), needs to be defined as well. It needs to be noted that, running the analytics pipeline on the Cloud Execution is the only available option if the data asset consumer selects the Apache Spark framework for running the analytics pipeline. In contrast, the data asset consumer is able to run the analytics pipeline on all the available execution locations (i.e., Cloud Execution, Server/Edge On-Premise Execution) using a Python environment. It needs to be noted that local execution can run analytics only over data that are stored locally at the moment (since no transfer of data from the cloud to the

On-Premise Environment is allowed). By clicking on the Save button, the Analytics Workbench page loads accordingly as described in the subsequent section.

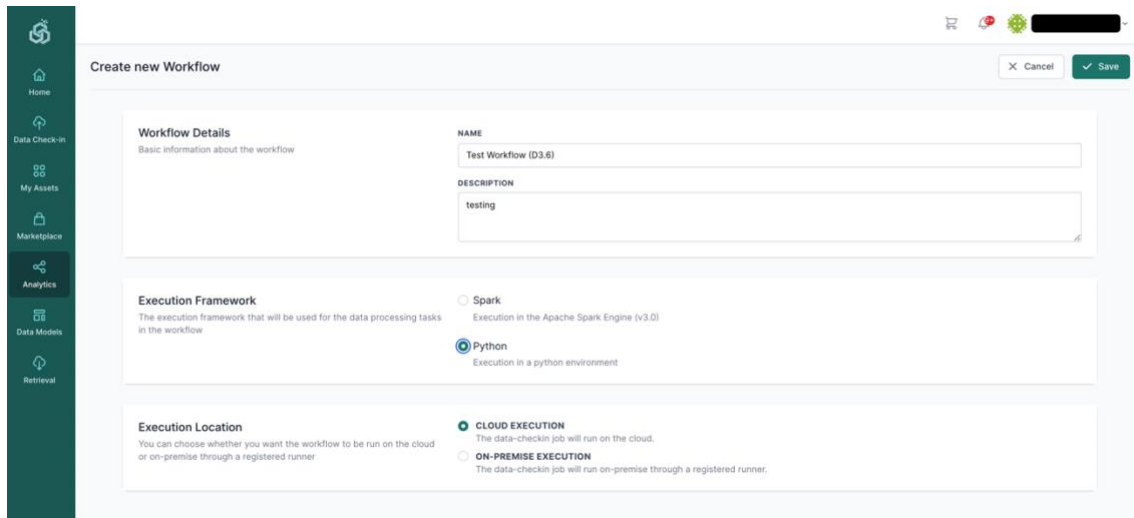


Figure 76: Create a new Analytics Pipeline

5.3 Configure an Analytics Pipeline

Through the Analytics Workbench, data asset consumers are able to appropriately configure a complete analytics pipeline by defining the necessary input data (input blocks), the data manipulation functions (data preparation), the Machine Learning (ML) or Deep Learning (DL) algorithms that are to be applied on the data, and the output data (output block) for storing the results of an analytics pipeline. This configuration is done by connecting these different pipes/function blocks such that the whole analytics pipeline is designed in a visual manner, in the form of easily configurable blocks that are appropriately connected to form the final data analysis pipeline. Towards the design of a complete data analytics pipeline, three main views are available to support visually the configuration, the validation of changes on the data sample after each step in the pipeline, and the configuration for visualising the results.

5.3.1 Graph View

The Graph view of the Analytics Workbench allows users to design (in a visual way) an analytics pipeline as shown in Figure 77.

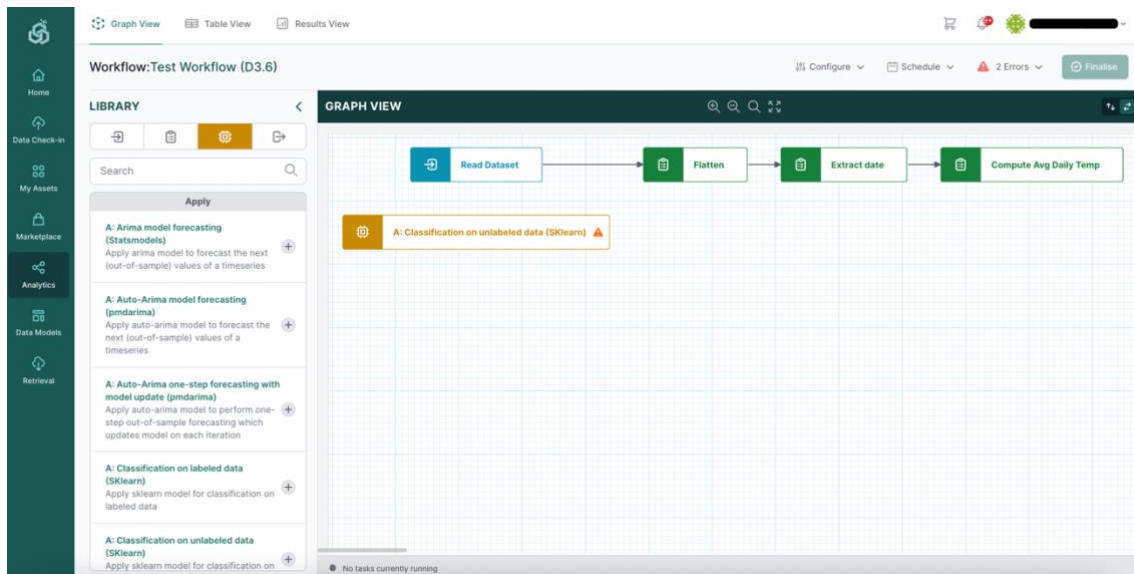


Figure 77: Configure an Analytics Pipeline – Graph View

By searching, browsing and filtering in the Library located at the left side of the view, the data asset consumers may find the relevant blocks for their pipeline, i.e. input blocks, data preparation blocks, machine learning (ML) / deep learning (DL) blocks, and output blocks. Upon locating a relevant block needed for the analytics pipeline in the Library, the data asset consumer can add them in the main Graph View. Adding a block, results in its appearance in the main Graph View without any connections (directed arrows between blocks), since the connection between the added blocks, is added when the data asset consumer configures the block connections in the right slideover.

Indicatively the data asset consumer should define the method for reading the input data, by adding a particular input block (e.g., Read Dataset which reads a stored data asset, and Read Result which reads an existing result from another data analysis pipeline that has been already executed), and inserting its associated parameters accordingly in the right slideover revealed after the block selection. Then, the data asset consumer may define data preparation rules by adding different processing functions (e.g., sort, drop null values, etc.), that are to be executed before the actual analytics functions. Once the data asset consumer has added the appropriate input and data preparation blocks, the relevant machine learning or deep learning blocks should be selected depending on the actual analysis that is intended to be done, as shown in Figure 78. The various machine learning and deep learning blocks are classified into three categories based on their use (e.g. evaluate, train, or apply), while the available machine learning algorithms (e.g., Binary Classification, Regression, Clustering, etc.) are allocated accordingly to these categories.



The libraries that are currently supported are: sklearn, MLlib, keras/tensorflow, pmdarima, statsmodels.

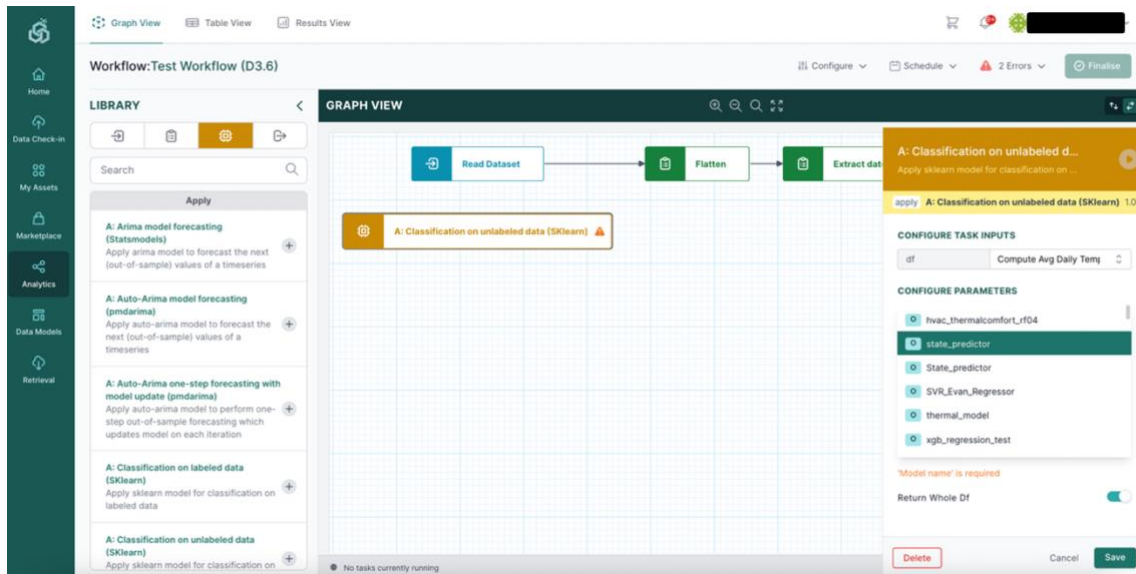


Figure 78: Analytics Pipeline - Machine Learning / Deep Learning Blocks

Finally, the data asset consumer should define the output block that denotes the way the data will be stored, or obtained from other mechanisms to create a result (e.g., through visualization, through the SYNERGY Platform’s APIs, etc.).

Practically, in the Library View, by selecting a block, the data asset provider may view and change the description of that block, see the type and version of the block, define the connections of the block (i.e., upstream tasks, downstream tasks), and define the parameters depending on the block type. By selecting the Save button, a quick run over the sample will be executed for the configured block in order to propagate the data structure to the next block and a relevant status icon will appear in the right side of the block in the Graph View. Once the blocks are configured correctly, a Valid icon will appear at the top bar of the Analytics Pipeline View. Additionally, the data asset provider may change the name or the description of the Analytics Pipeline by selecting the Configure button.

The available actions in the blocks include: (a) test run in order to execute the pipeline up to the specific block for a small data sample, (b) settings in order to edit or delete a block from the pipeline. In case there are any validation errors in the pipeline, they will be displayed in the appropriate blocks, and a summary of the validation results will be displayed in the Validation Results as shown in Figure 79. The data asset provider may proceed to view the results of the

test run by selecting the Table View (Section 5.3.2) from the Workflow Designer top bar, and follow the workflow as described in the subsequent section.

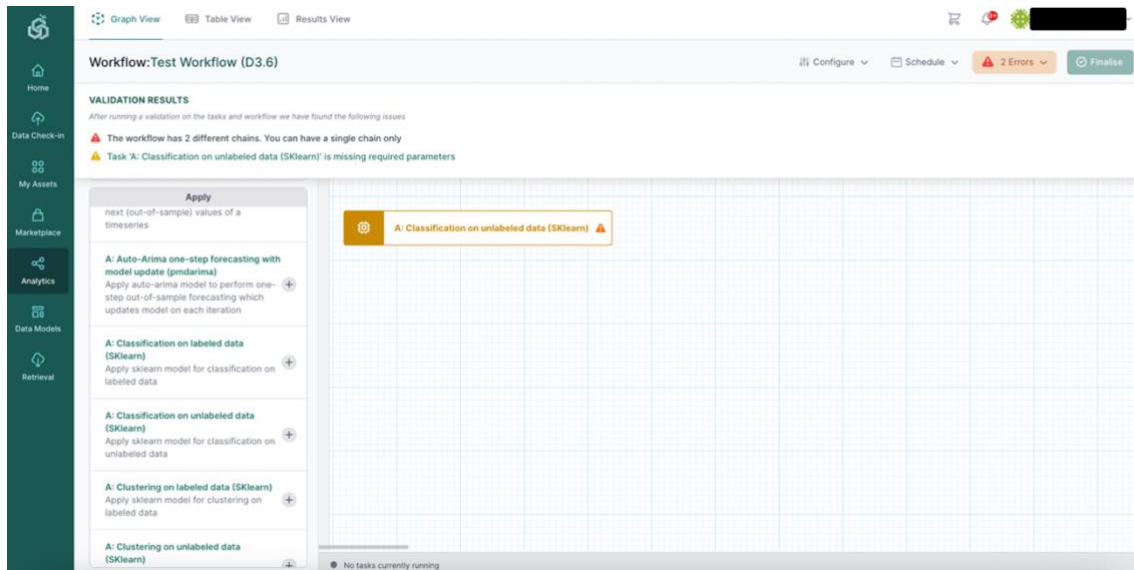
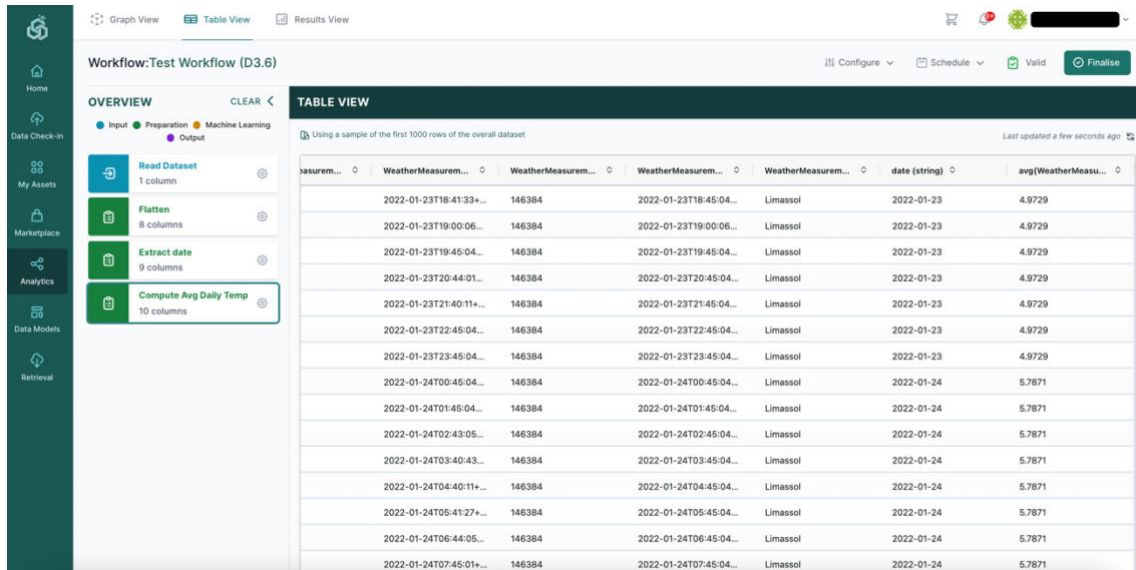


Figure 79: Analytics Pipeline - Validation Results

5.3.2 Table View

In the Table view, depicted in Figure 80, data asset consumers obtain an overview of a data sample, and how data change each time they are processed according to the blocks defined in the previous step. In particular, the data asset provider may view the blocks that constitute the pipeline in the left slideover, while by selecting a particular block the corresponding workflow is highlighted, and the results of the test run appear in the main Table View. Additionally, by selecting a particular block, the data asset consumer may change the settings that were defined previously in the Graph View. In the main Table View, the data asset consumer may view the different columns of the outcomes and may sort them, search for a specific value or resize them.



WeatherMeasure...	WeatherMeasure...	WeatherMeasure...	WeatherMeasure...	date (string)	avg(WeatherMeasu...
2022-01-23T18:41:33+...	146384	2022-01-23T18:45:04...	Limassol	2022-01-23	4.9729
2022-01-23T19:00:06...	146384	2022-01-23T19:00:06...	Limassol	2022-01-23	4.9729
2022-01-23T19:45:04...	146384	2022-01-23T19:45:04...	Limassol	2022-01-23	4.9729
2022-01-23T20:44:01...	146384	2022-01-23T20:45:04...	Limassol	2022-01-23	4.9729
2022-01-23T21:40:11...	146384	2022-01-23T21:45:04...	Limassol	2022-01-23	4.9729
2022-01-23T22:45:04...	146384	2022-01-23T22:45:04...	Limassol	2022-01-23	4.9729
2022-01-23T23:45:04...	146384	2022-01-23T23:45:04...	Limassol	2022-01-23	4.9729
2022-01-24T00:45:04...	146384	2022-01-24T00:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T01:45:04...	146384	2022-01-24T01:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T02:43:05...	146384	2022-01-24T02:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T03:40:43...	146384	2022-01-24T03:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T04:40:11...	146384	2022-01-24T04:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T05:41:27+	146384	2022-01-24T05:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T06:44:05...	146384	2022-01-24T06:45:04...	Limassol	2022-01-24	5.7871
2022-01-24T07:45:01...	146384	2022-01-24T07:45:04...	Limassol	2022-01-24	5.7871

Figure 80: Configure an Analytics Pipeline – Table View

5.3.3 Results View

Once the data asset consumer is satisfied with the defined and configured data analytics pipeline, he/she may configure the desired output / visualization of the results from the Results View as shown in Figure 81. In particular, the data asset consumer may select the output block that he/she is interested to visualize, and then the type of diagram that needs to be visualized from the left slideover menu that includes all the available visualization types. This menu provides several types of diagrams for visualization such as line graph, scatter plot, and bar chart. Following that, the data asset consumer needs to select the appropriate output blocks in order to data Models that are to be visualized, from the Configuration right slideover menu. The parameter selection needs to be done for each axis according to the type of the diagram, and the needs of the data asset consumers with regards to what is to be visualized. In addition, the data asset consumer may add a title for the diagram, legends for the data, axes labels for the diagram, as well as to select among various other options such as to add grids, and tooltip to the diagram.

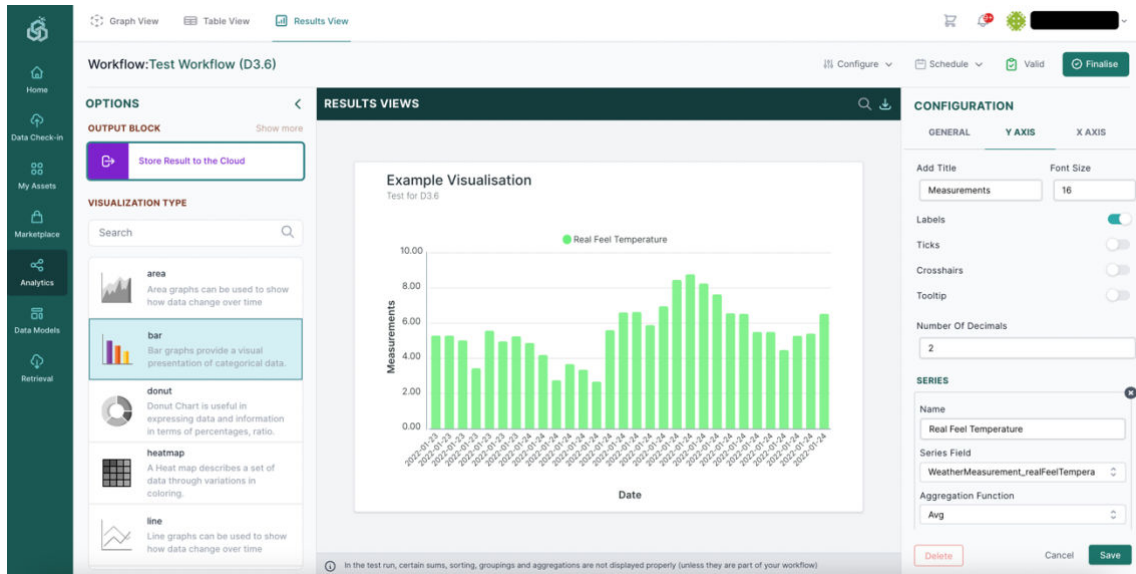


Figure 81: Configure an Analytics Pipeline – Results View

5.3.4 Schedule Execution

Apart from the configuration of the different input, processing, analytics and output blocks that need to be added in a complete analytics pipeline, the data asset consumer may also configure the scheduling of the pipeline execution, by selecting the Schedule option from the Analytics Workbench bar, as shown in Figure 82. A top slideover to define a schedule (or multiple schedules) is displayed allowing the data asset provider to select the execution period (i.e., start date, end date), the frequency of the execution (i.e., hourly, daily, weekly, monthly), and the exact time of execution depending on the frequency.

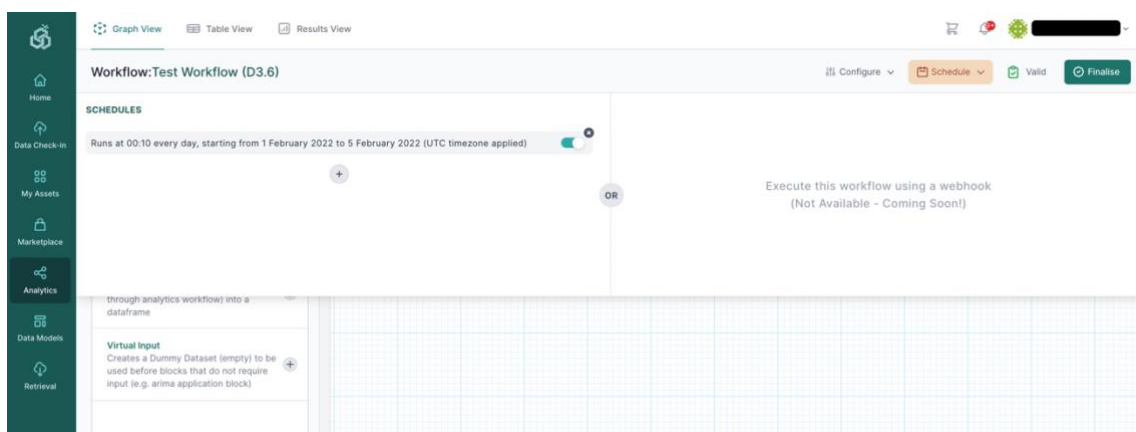


Figure 82: Configure an Analytics Pipeline – Schedule Execution

5.3.5 Execution History

By selecting the Execution History option for a particular Analytics Pipeline, the execution logs will appear as depicted in Figure 83. Within this view, the data asset consumer can see the execution summary information such as the number of total executions, the number of successful executions, the number of failed executions, and the average execution time. Below these, the data asset consumer may view more details regarding each execution including its timestamp, status, and available actions. When an execution is queued, there is only one available option that is to cancel the next execution as the schedule shall continue (skipping the next execution only). When an execution has successfully been completed or has failed, the data available options are: a) to view the details which opens a table that includes the task details, b) to visualize the results in the Results View. It needs to be noted that the execution logs are organized in a listed view that can be sorted by the date of execution or status accordingly.

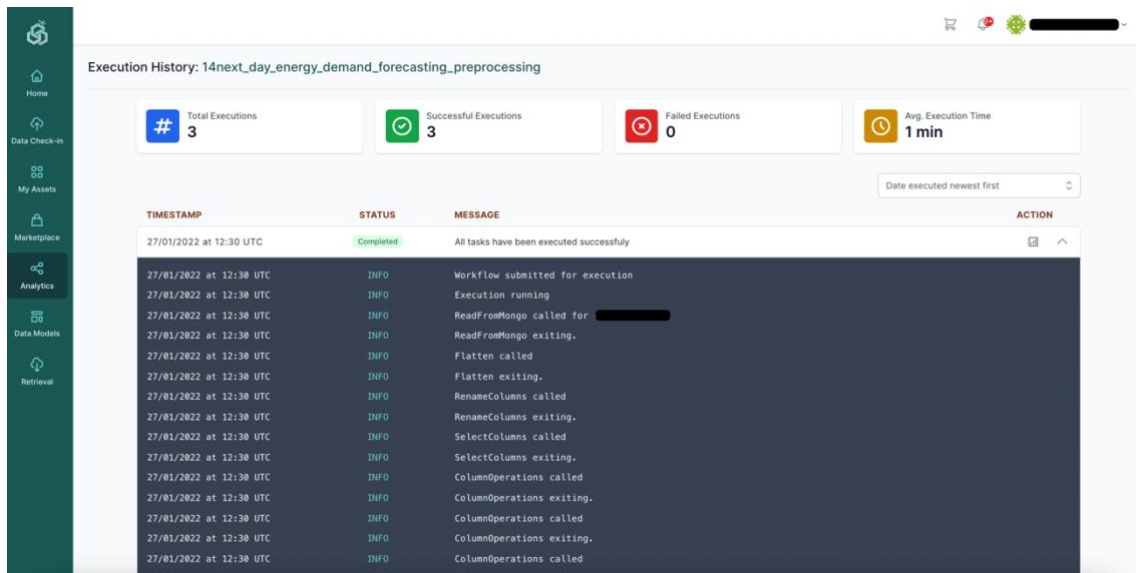


Figure 83: Analytics Pipeline - Execution History

5.4 Register a pre-trained Analytics Model

Although models can be trained and applied in the analytics pipelines through the respective machine learning and deep learning blocks (as described in section 5.3.1), they need to be registered in the SYNERGY Platform in order to be considered as available and to be shared with other stakeholders of the electricity data value chain (outside an organisation). By opting to register a model (from Figure 75), a user needs to provide the model profile (e.g. if it has been trained within the platform or externally to the platform, the library it has been used, etc.) and

the licensing information. Since directly writing code is not allowed in the SYNERGY Platform due to security reasons, a data asset consumer is able to register pre-trained analytics models that have been created “offline”, i.e. outside the SYNERGY Platform, by uploading the relevant model file(s), the sample data and defining and ordering the model features as depicted in Figure 84.

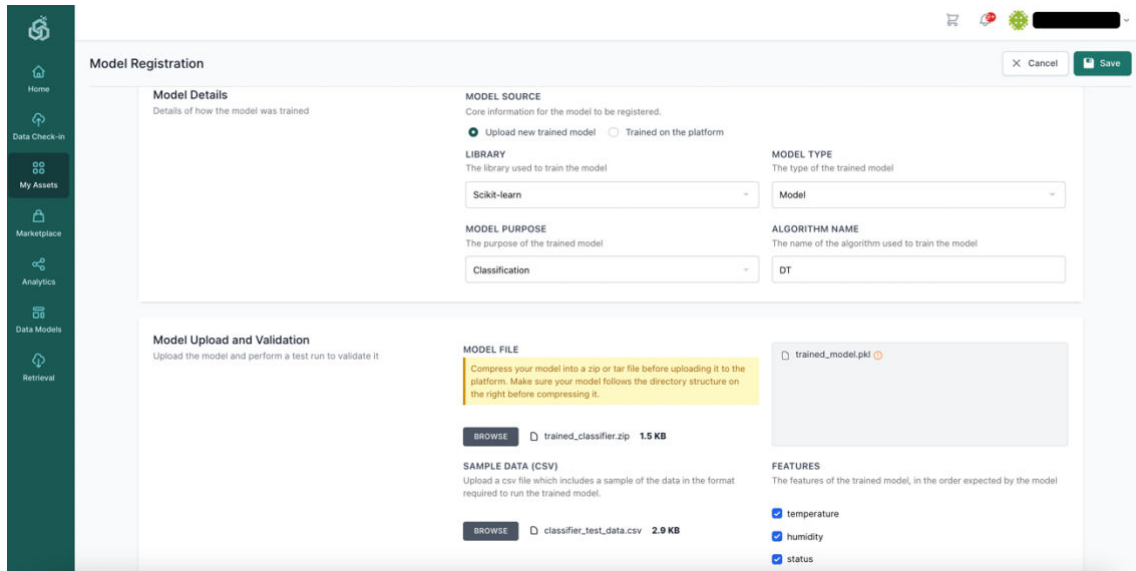


Figure 84: Register a pre-trained Model

Once the user saves the model details, the SYNERGY Platform validates the model by applying it to the sample data provided as depicted in the following figure.

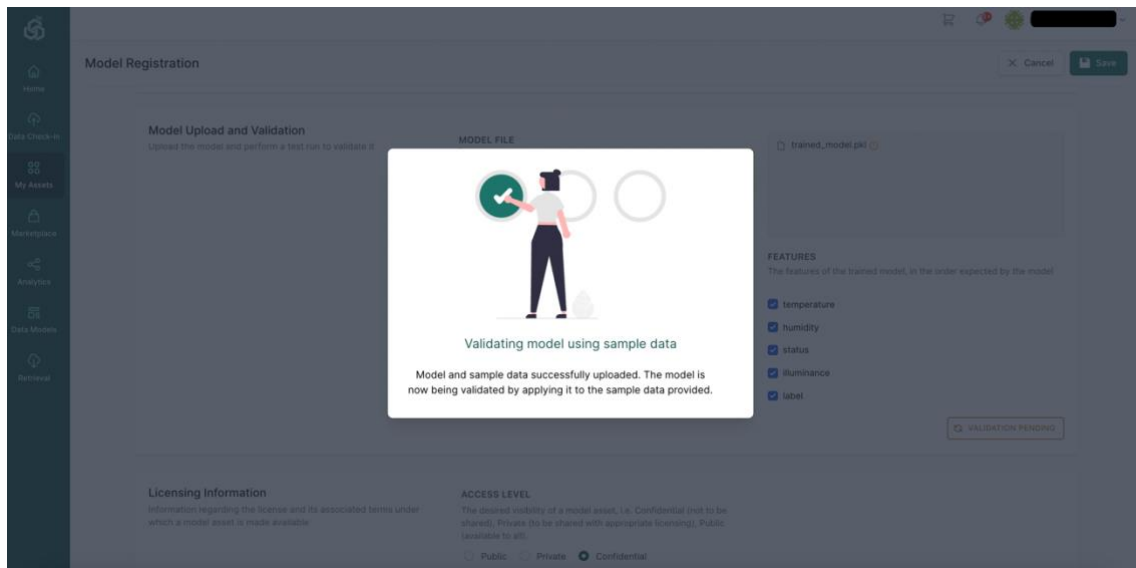


Figure 85: Validating a pre-trained Model using sample data

If the user has properly packaged the model (e.g. compatible library version) and provided all the necessary info for the model (features), the model validation succeeds (as depicted in the following figure), otherwise he/she gets an error message.

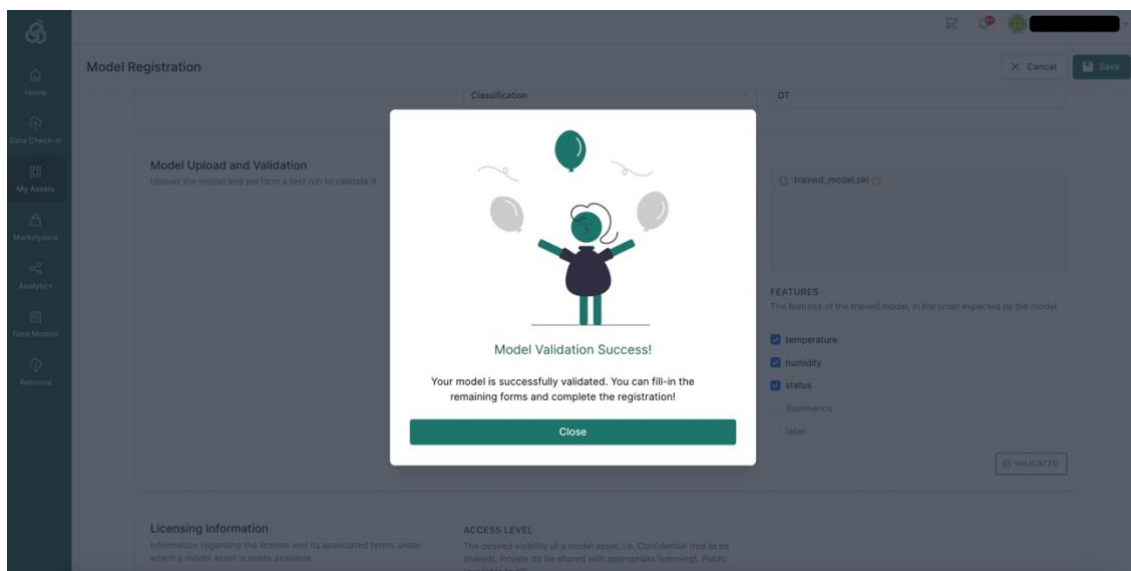


Figure 86: Successful validation of a pre-trained Model

5.5 Make a pre-trained Analytics Model or the Results of an Analytics Pipeline, available in the SYNERGY Marketplace

In order to make a derivative asset (i.e. a pre-trained analytics model or the results of an analytics pipeline) available to the SYNERGY Marketplace, a two-step process needs to be followed: (a) the relevant metadata need to be appropriately filled in, and (b) the contributing data asset providers to the specific derivative asset need to provide their explicit consent through derivation contracts.

Through the menu “My Assets” - Models, the users can access the models that belong to their organisation and are able to edit their metadata, including the general information, the licensing details and the access policies, as depicted in the following figure.

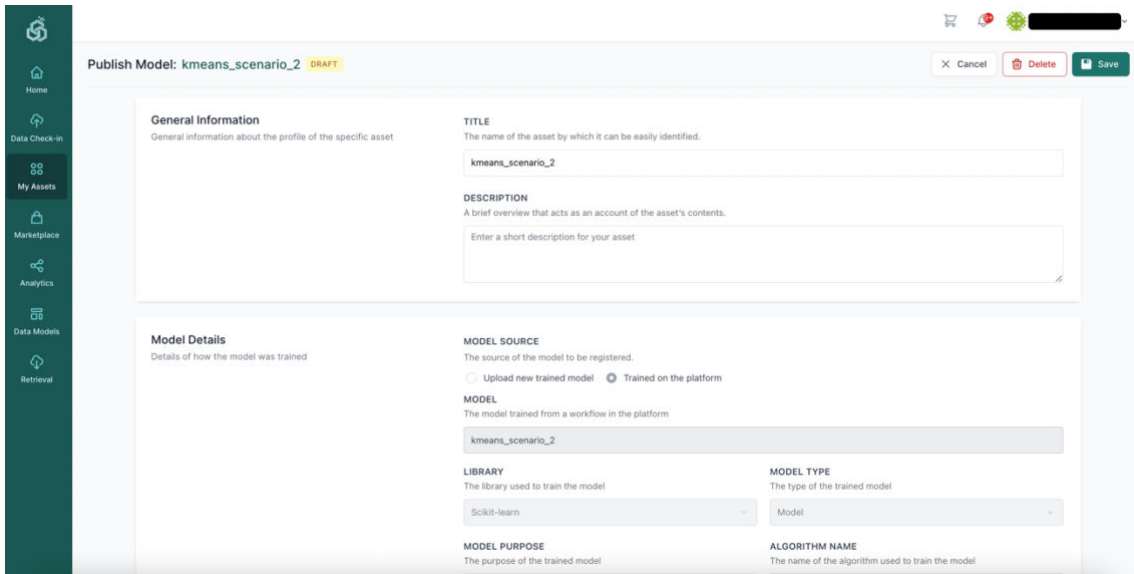


Figure 87: Provide metadata info for a pre-trained model

If the derivative asset is a result, similarly through the menu “My Assets” - Results, the users can access the results that belong to their organisation and are able to edit their metadata, including the general information, the distribution details, the licensing details and the access policies, as depicted in the following figure.

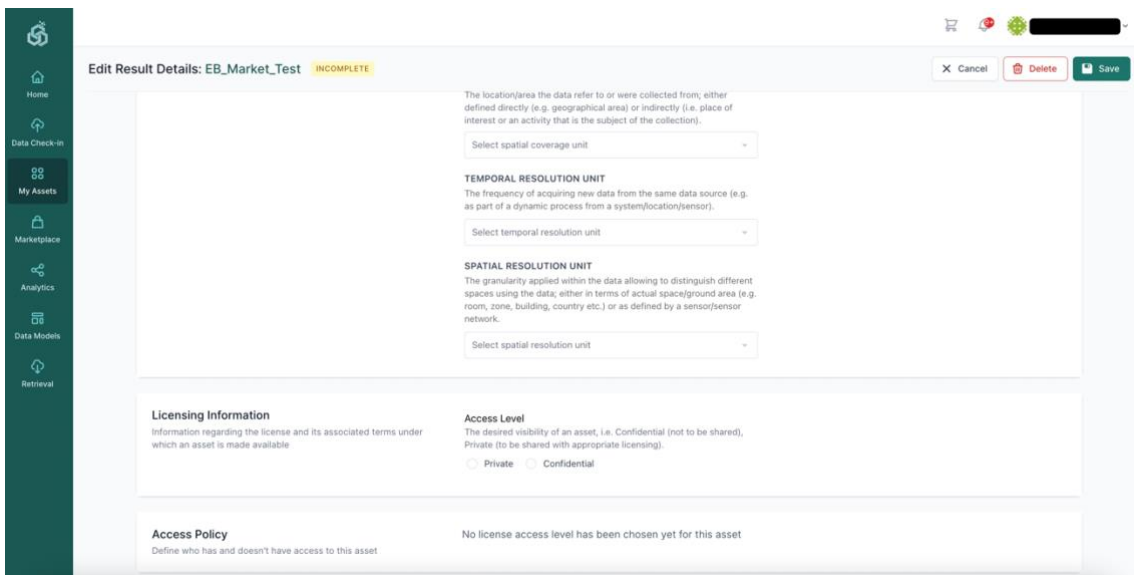


Figure 88: Provide metadata info for an analytics pipeline results

In case that data assets from other organisations have been involved in the creation of a pre-trained model or a result (e.g. providing input datasets and/or pre-trained models, respectively),

then the user needs to explicitly add the model or result to the SYNERGY Marketplace as depicted in the following figure.

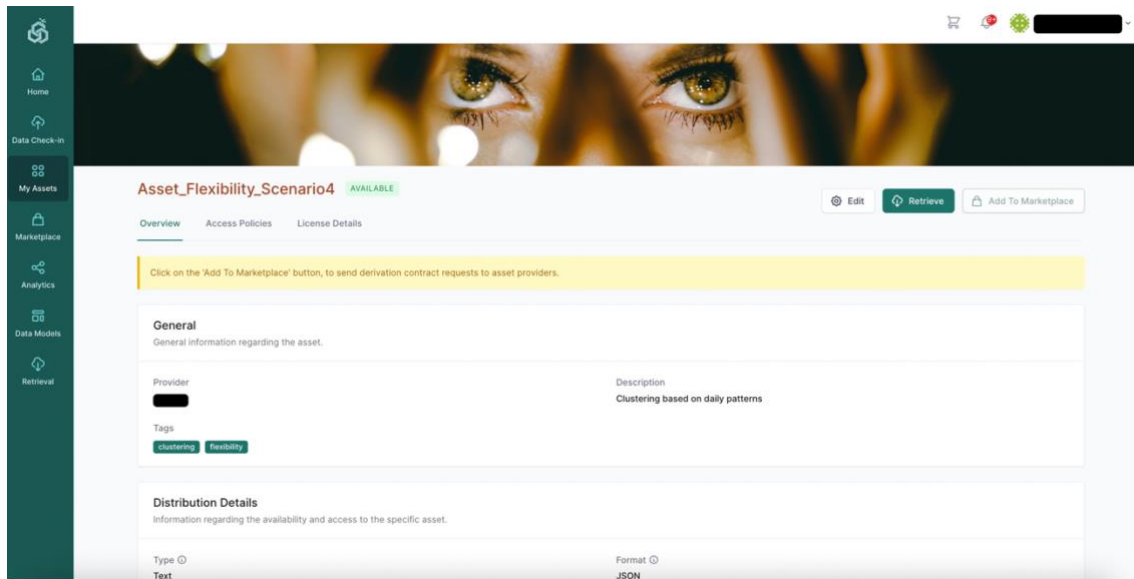


Figure 89: Result profile page, allowing to add it to the Marketplace

Opting to add a derivative asset to the SYNERGY Marketplace means that appropriate derivation contracts are sent to the involved data providers (with whom active acquisition contracts are in place). Each involved data asset provider may navigate to the derivation contract request (as depicted in the following figure) and decide whether to offer a derivation contract or not. Declining to offer a derivation contract means that the respective derivative data asset (pre-trained model or result) cannot become available in the SYNERGY Marketplace, yet the data asset consumer can continue utilising it as before (based on the active acquisition contract’s terms).

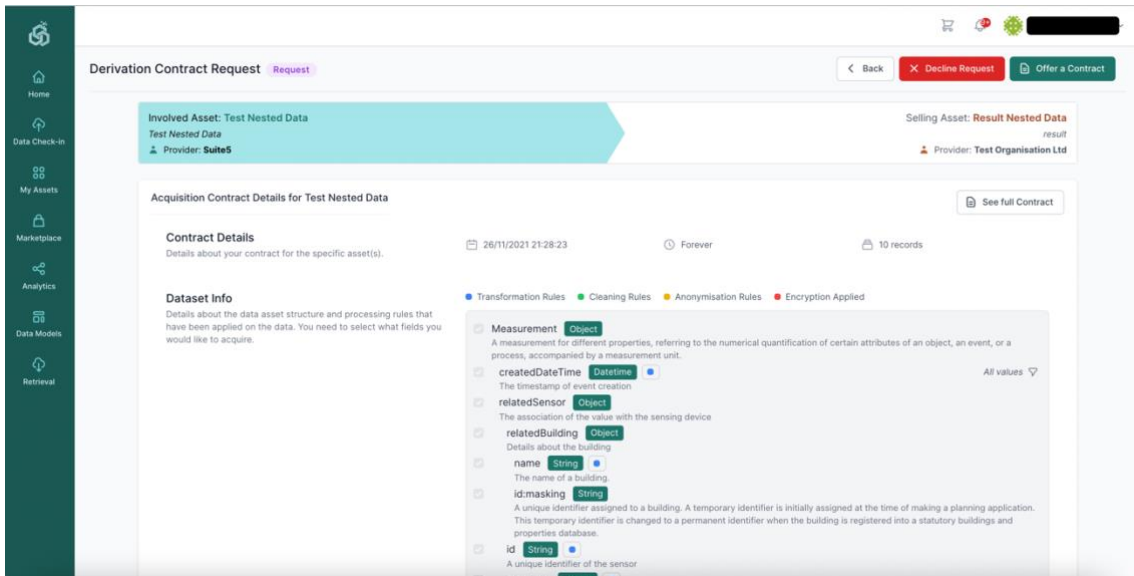


Figure 90: Derivation Contract Request to determine the availability of a derivative asset in the SYNERGY Marketplace

The involved data asset(s) provider(s) can edit the licensing metadata with the exception of metadata that are directly related to the indirect availability of the asset in the SYNERGY Marketplace (as depicted in the following figure). Once the data asset(s) provider(s) sign the derivative contracts (in the SYNERGY blockchain), the derivative asset becomes visible in the marketplace.

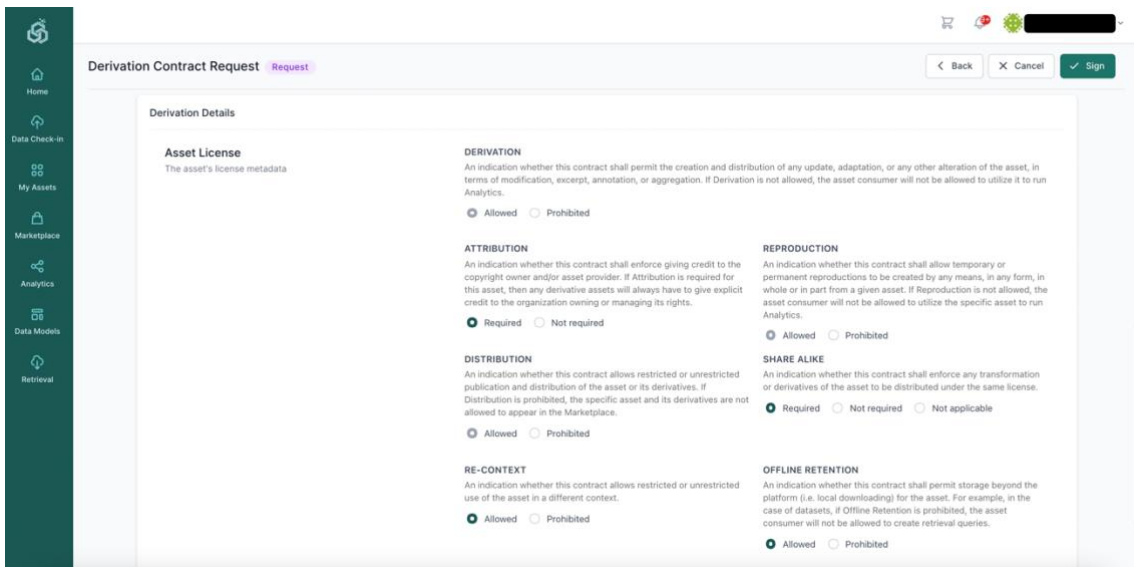


Figure 91: Derivation Contract Preparation (Data Asset Provider Perspective)

At any moment, users can view all derivation contracts in which assets of their organisations are involved as depicted in the following figure.



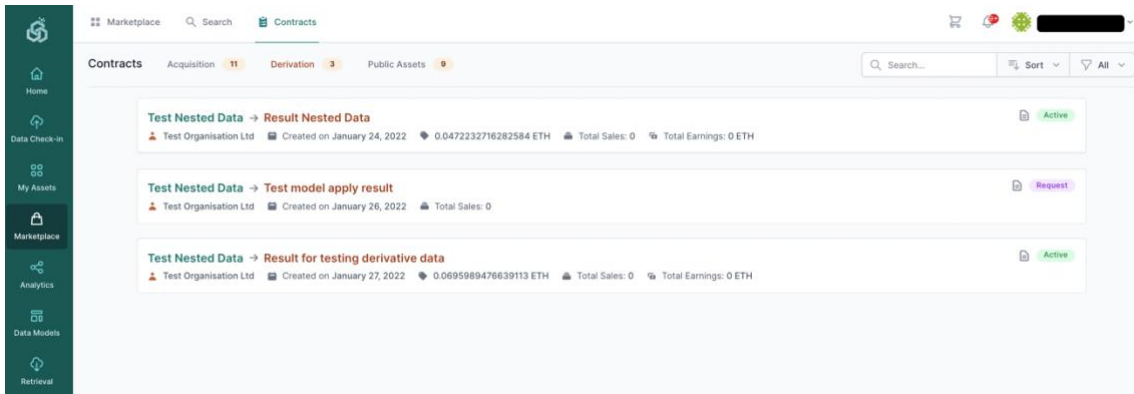


Figure 92: View the Derivation Contracts List

5.6 Visualize the Results of an Analytics Pipeline

Once an analytics pipeline has been finalised and has been successfully executed, the data asset consumer may view the results as shown in Figure 93. The results are adapted based on the configuration of the visualization, as defined in the Results View page described in Section 5.3.3. Moreover, the user is able to download the visualisation locally, or to retrieve the results from a customised API endpoint (as described in section 4.4).



Figure 93: Analytics Pipeline – View Results of an Analytics Pipeline

6 Additional Platform Functionalities

In addition to the functionalities that were described in the former sections, the SYNERGY Platform provides some additional functionalities to different types of users (i.e., platform’s administrators, data asset consumers, data asset providers, and others) enabling platform support functions. Such support functionalities include: (a) Common Information Model Lifecycle Management for the CIM administrators, (b) Edit Organization Profile, (c) Edit User Profile, and (d) Wallet Management.

6.1 CIM Manager

The Platform’s model administrators are allowed to create, edit, update and deprecate data models representing data that are exchanged between the stakeholders of the energy data value chain, in accordance with the CIM lifecycle management approach that was defined in the SYNERGY Deliverable D3.1. This functionality is offered by the CIM Manager component, and can be accessed through the Models view.

Within the CIM Manager, depicted in Figure 94, the model manager is able to view: (a) the list of concepts that are included in the data model at the left side of the page, (b) the fields/concepts that are included (or linked) in the selected concept at the middle of the page, and (c) more details about the selected field such as its description, its mapping to standards, and its related terms, at the right side of the page.



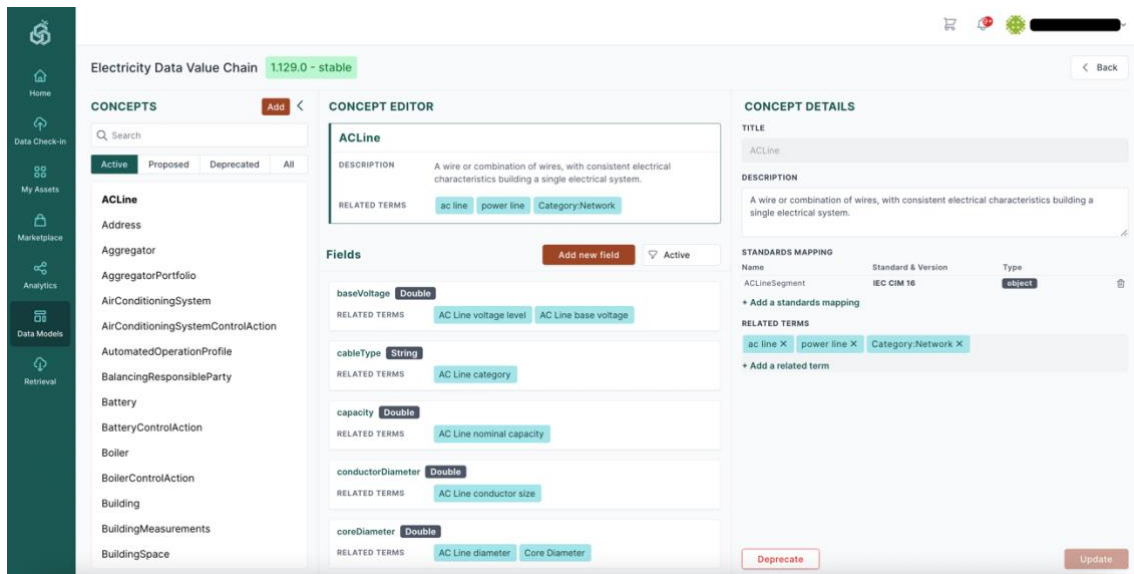


Figure 94: Common Information Model (CIM) Manager – Concepts/Fields View

In this page, the model manager may add, edit, or delete existing concepts or fields according to the stakeholders' needs. In order to maintain the CIM up-to-date according to the latest needs of the data asset providers in the SYNERGY Platform, the model manager is able to edit the concepts and fields of the selected data model, which generates a new major or minor version of the CIM (depending on the evolution rules that are triggered under the hood) and makes the new data model and its associated concepts/fields available in the Mapping configuration stage. It is worth mentioning that the model manager is notified (in the platform and/or via email) about any proposed concepts by the data asset providers in order to timely take appropriate actions.

In particular, the model manager may add a concept to the CIM by selecting the Add button in the Concepts section to enable the right Concept Creator section where he/she needs to insert the title, the description, the mapping to a particular standard (if any), as well as to add related terms for the concept. Additionally, the model manager can select to import an existing concept from another data model by selecting the Import from another Data Model option (e.g. from a deprecated CIM version) as shown in Figure 95. Finally, the concept is created by selecting the Create button which enables the Concept Editor section in order for the model manager to provide the fields to be included in the concept.

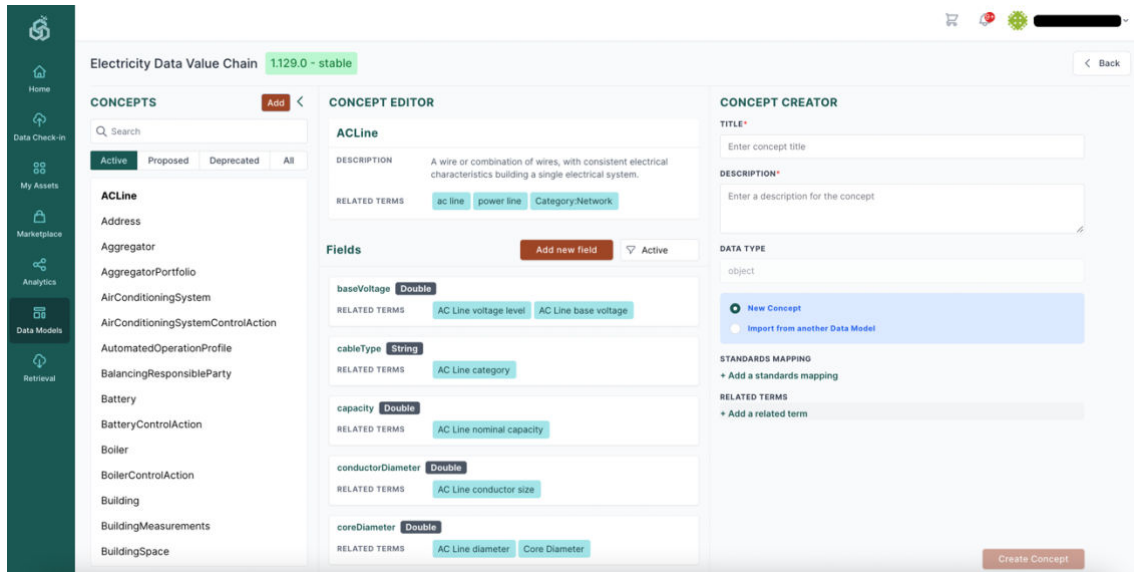


Figure 95: Common Information Model (CIM) Manager – Concept Creator

Once the concept is created, the model manager can add the fields or sub-concepts, by selecting the “Add new Fields” option in the Concept Editor section, which will enable the Field Creator section at the right side of the page as Figure 96 depicts. In this section, the model manager can provide the title and description for the field, select its data type (e.g. string, double, integer, object, etc), add a standard mapping (referring to the fields of certain standards), add related terms, and add metadata to the field. Appropriate metadata can be added to the field depending on the data type and including information regarding different aspects of the data that could be potentially uploaded to the particular field.

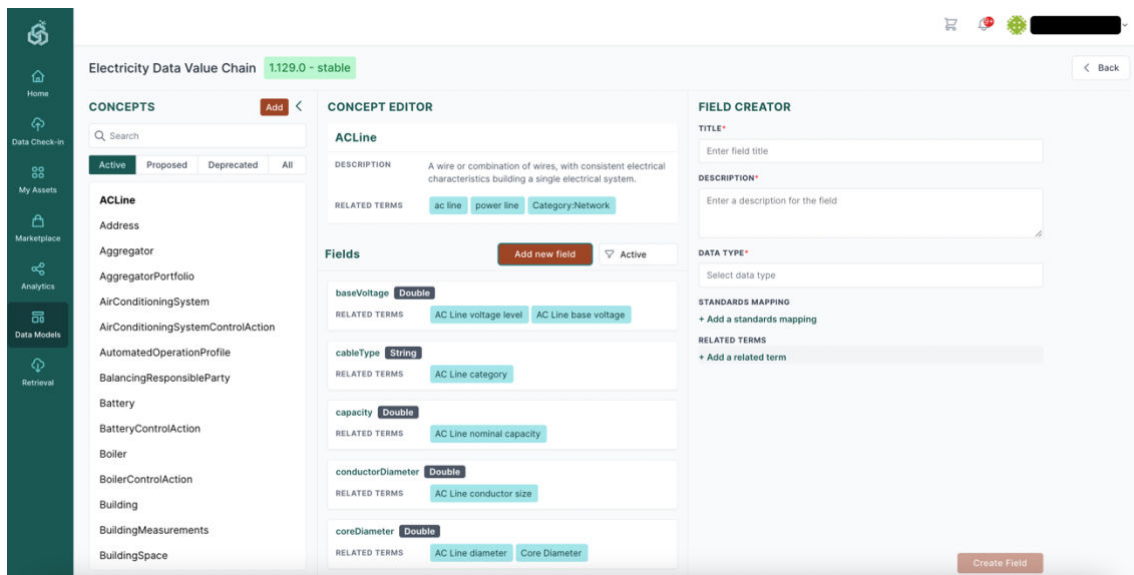


Figure 96: Common Information Model (CIM) Manager – Field Creator

6.2 Manage an Organization Profile

The organisation’s manager (who is also the organisation’s legal representative, eligible to sign any data asset contracts in the SYNERGY Platform) may edit the organisation’s profile, by navigating to the Edit Organisation Profile page as shown in Figure 97. The organisation’s manager may edit the organisation’s business name, by which it is typically identified. In addition, the manager can add a description that provides a brief overview regarding the organisation’s activities. The organisation type that distinguishes the organisation within the electricity data value chain, needs to be selected as well (even though this functionality is deactivated upon the initial organization registration to avoid potential misuse). The organisation’s manager may add, edit, or remove a department (including its name, address, city, and country) to the Departments List as well.

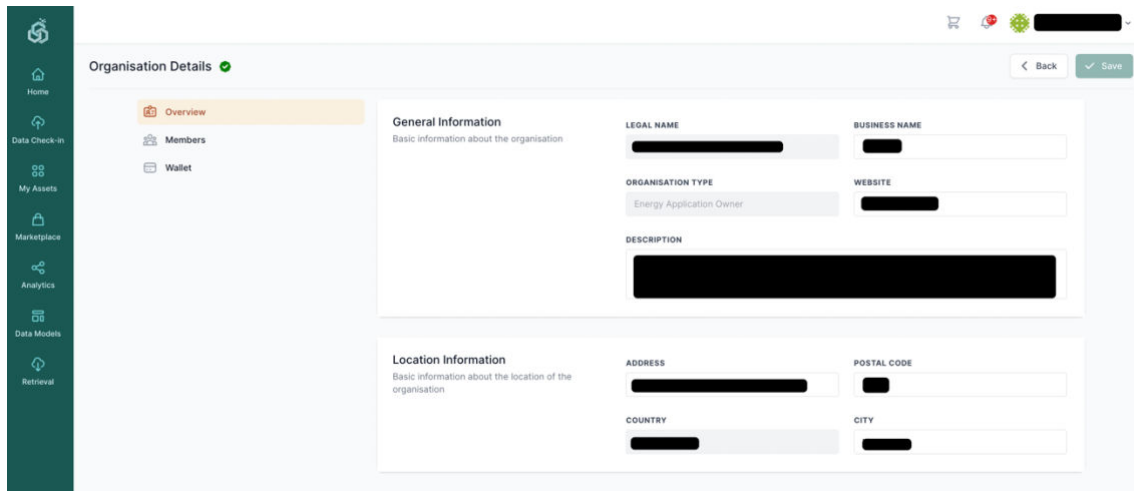


Figure 97: Edit Organisation Profile - Overview

In addition, users can be added in the Users List by defining the department they work, their role in the organisation, and their status (i.e., Active, Invited, Blocked), while a user can be activated (if Blocked) or deactivated (if Invited or Active) by the manager by selecting the corresponding buttons, as depicted in Figure 98.

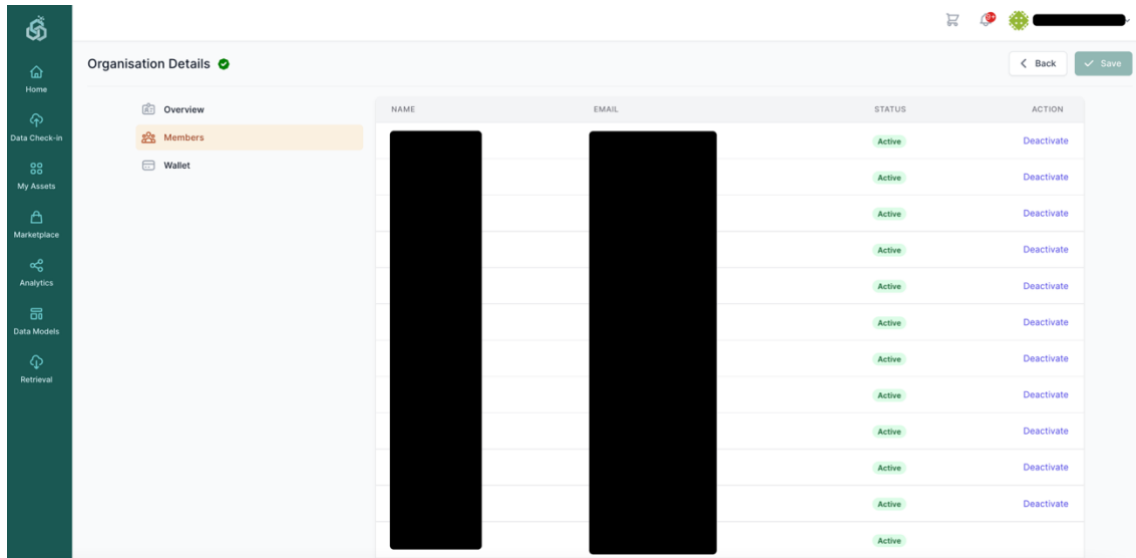


Figure 98: Edit Organisation Profile - Users List

6.3 Manage the User Profile

Users may edit their individual user profiles, by navigating to the Personal Profile page that is loaded by selecting the name of the user at the top right side of the SYNERGY Platform. As depicted in Figure 99, the user may update basic information such as the first and last name, as well as to change their current password to a new one, by filling the corresponding text boxes.

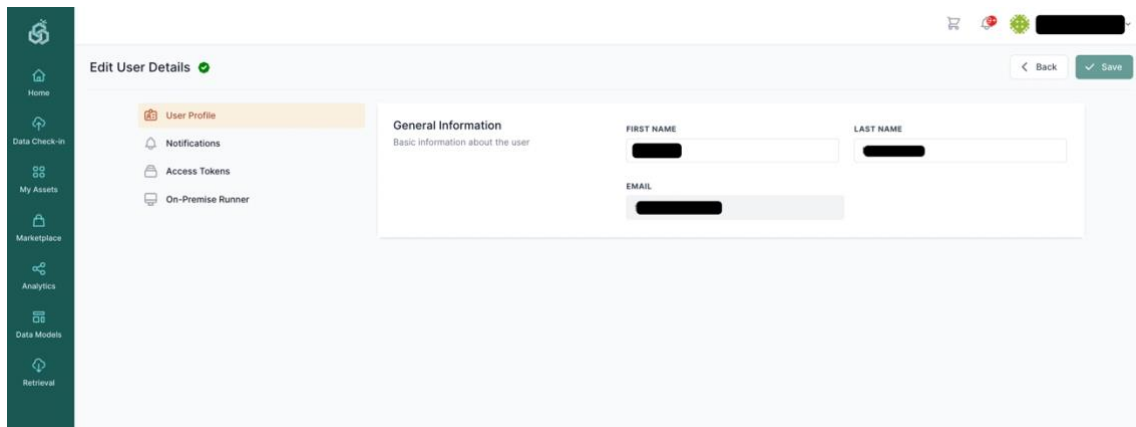


Figure 99: Edit User Profile – Overview

6.3.1 Configure the Notifications settings

In the Notifications tab, a user is able to configure what notifications he/she would like to receive in the SYNERGY Platform as well as via email. In the case of email notifications, the user can opt

for daily digest emails instead of instant notifications for any notification event of his/her preference.

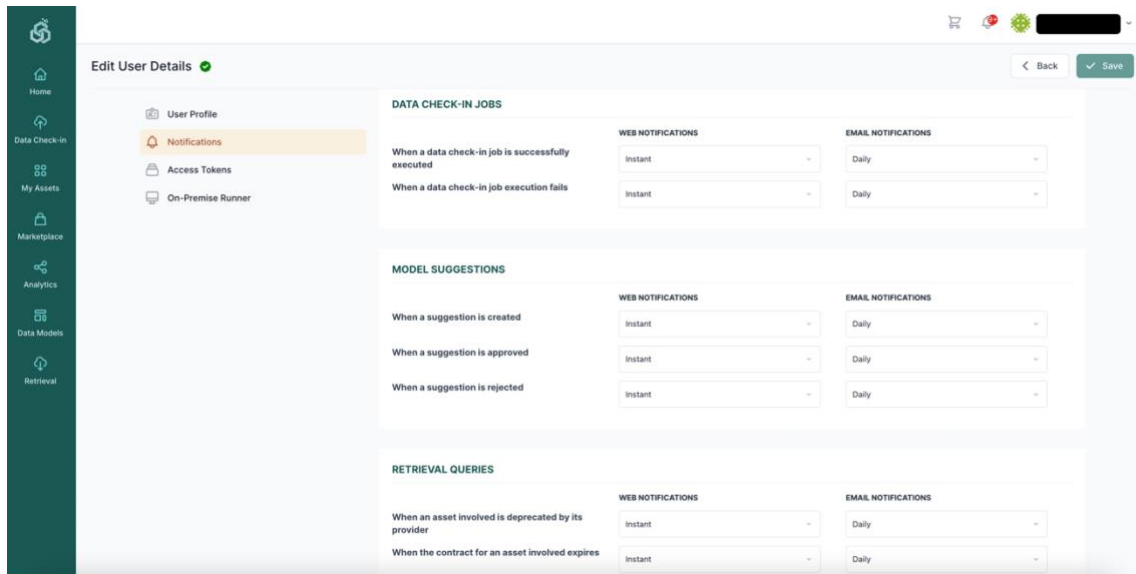


Figure 100: Edit User Profile – Notifications

6.3.2 Generate Access Tokens

By selecting the Access Tokens tab, the users are able to view their tokens (for retrieving or uploading data to the SYNERGY Platform through the SYNERGY APIs). By selecting the Generate new token, the user is asked to provide a name for the token that is to be generated and select the scope of the intended use that is to retrieve data from the platform’s API, or to upload data to the platform’s API, or both. Finally, the user is able to delete generated personal access tokens by selecting the Delete button located at the right side of each access token in the Access Tokens tab, as depicted in Figure 101.

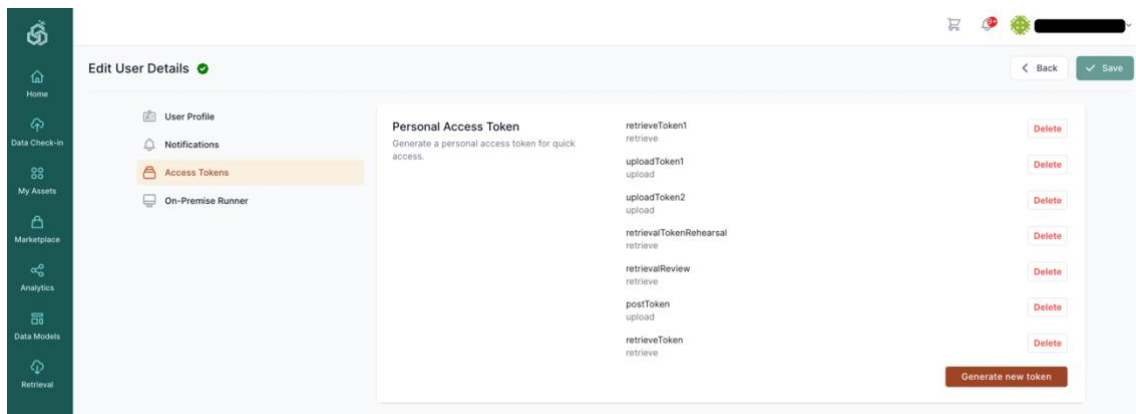


Figure 101: Edit User Profile - Access Tokens – Generate Token

6.3.3 Register an On-Premise Execution Environment

The user may download the On-Premise Execution Environment tab as shown in Figure 102. In particular, the user needs to select the operating system of his/her physical machine in order to start downloading the appropriate version of the Server and Edge On-Premise Environments that are available in a tray and headless release for Windows, MacOS, Linux and Edge.

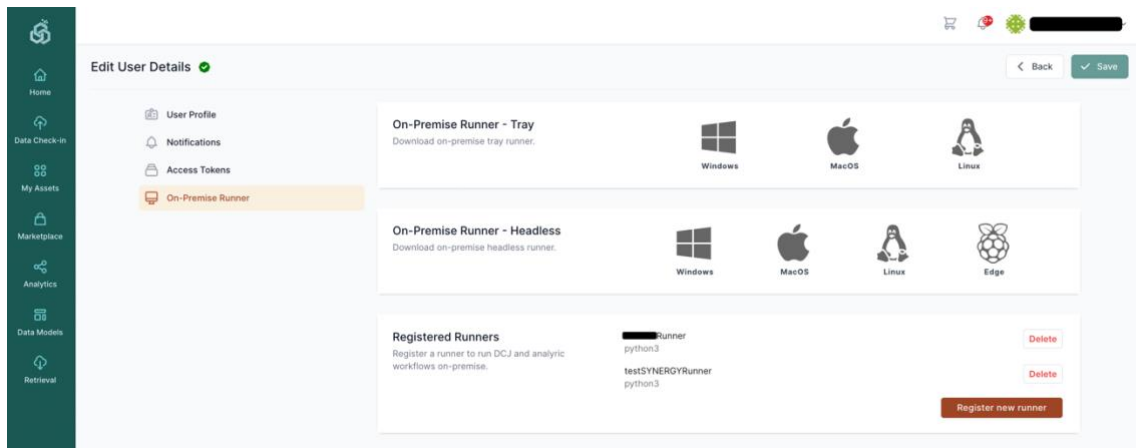


Figure 102: Edit User Profile - On-Premise Execution Environment

However, a prerequisite step to execute the On-Premise Execution Environment is to register it by selecting the “Register a new runner” button which then enables a text box asking to provide a name for the Execution Environment. As soon as the name is inserted correctly, the corresponding URL and code for the registered On-Premise Execution Environment will appear allowing the execution of the Tray On-Premise Execution Environment, as shown in Figure 103. The generated URL and code need to be inserted during the installation phase on the On-Premise Execution Environment on the user’s physical machine.

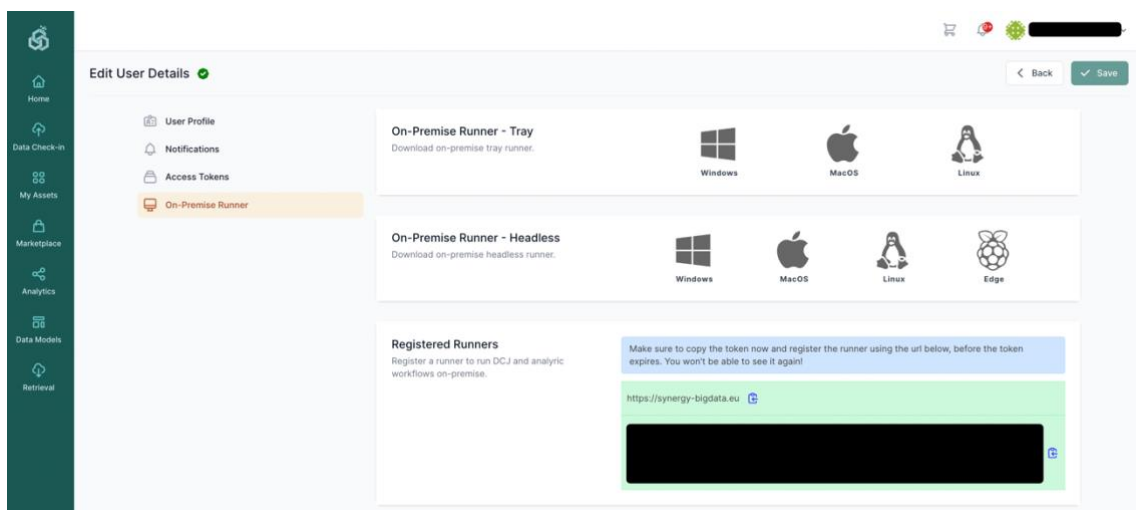


Figure 103: Edit User Profile - On-Premise Execution Environment - Register a New Runner

6.4 Manage the Organization’s Wallet

An organization’s manager may view the contracts of his/her organization by navigating to the Contracts view of the Marketplace, as shown in Figure 55. In the case that the organization does not have a linked wallet, the manager may import an existing wallet, or create a new one as described in the subsequent sections (Section 6.4.1 and Section 6.4.2, respectively). It needs to be noted that, if data asset providers have already initiated requests for data assets (that do not belong to their organization) but the organization manager has not set up a wallet yet, they will not be able to proceed in buying a data asset, until a wallet is generated in the Contracts tab.

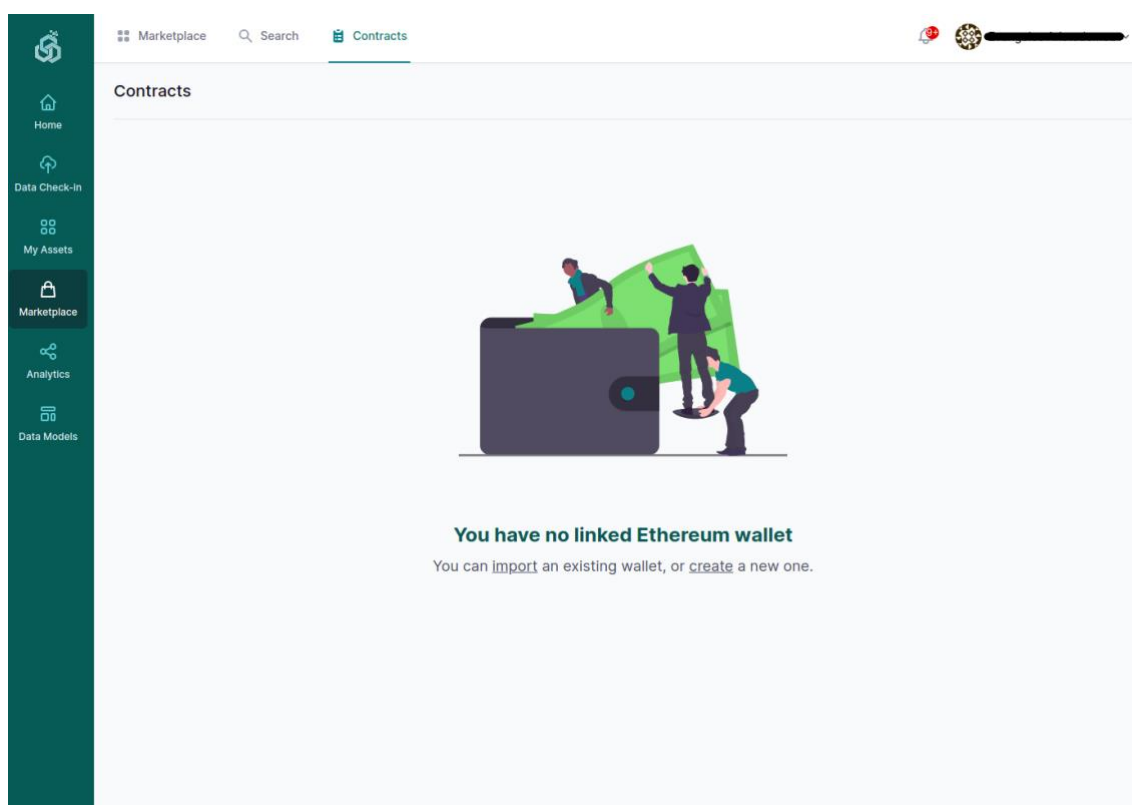


Figure 104: Contracts List in case there is no linked Ethereum Wallet

6.4.1 Import an Existing Wallet

The organization’s manager may import an existing wallet by entering the private key and the wallet password, as depicted in Figure 105.

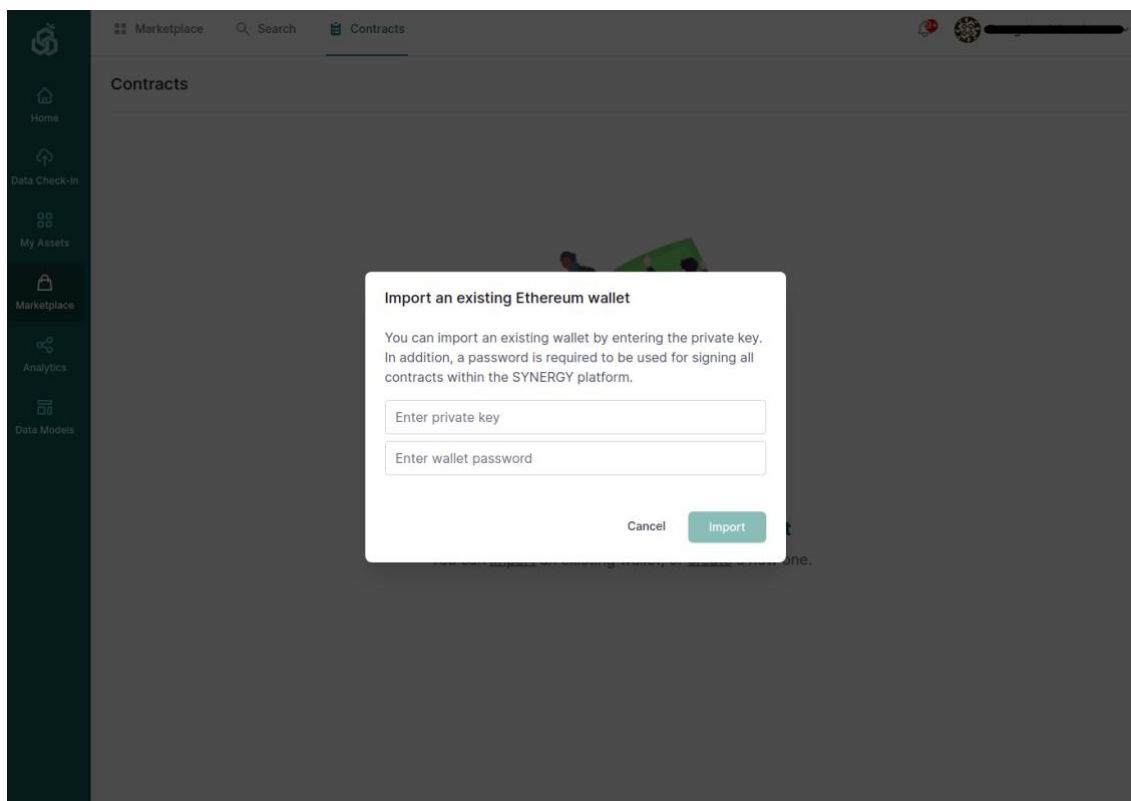


Figure 105: Import an Existing Wallet

6.4.2 Create a New Wallet

An organization that does not have a wallet can create a new one in the Contracts view of the Marketplace. A wallet password needs to be entered accordingly to encrypt the organization’s wallet, as Figure 106 depicts. It needs to be noted that this password will be used for signing all contracts within the SYNERGY Platform. By selecting the Create button, the platform starts generating and encrypting the wallet as Figure 107 shows.

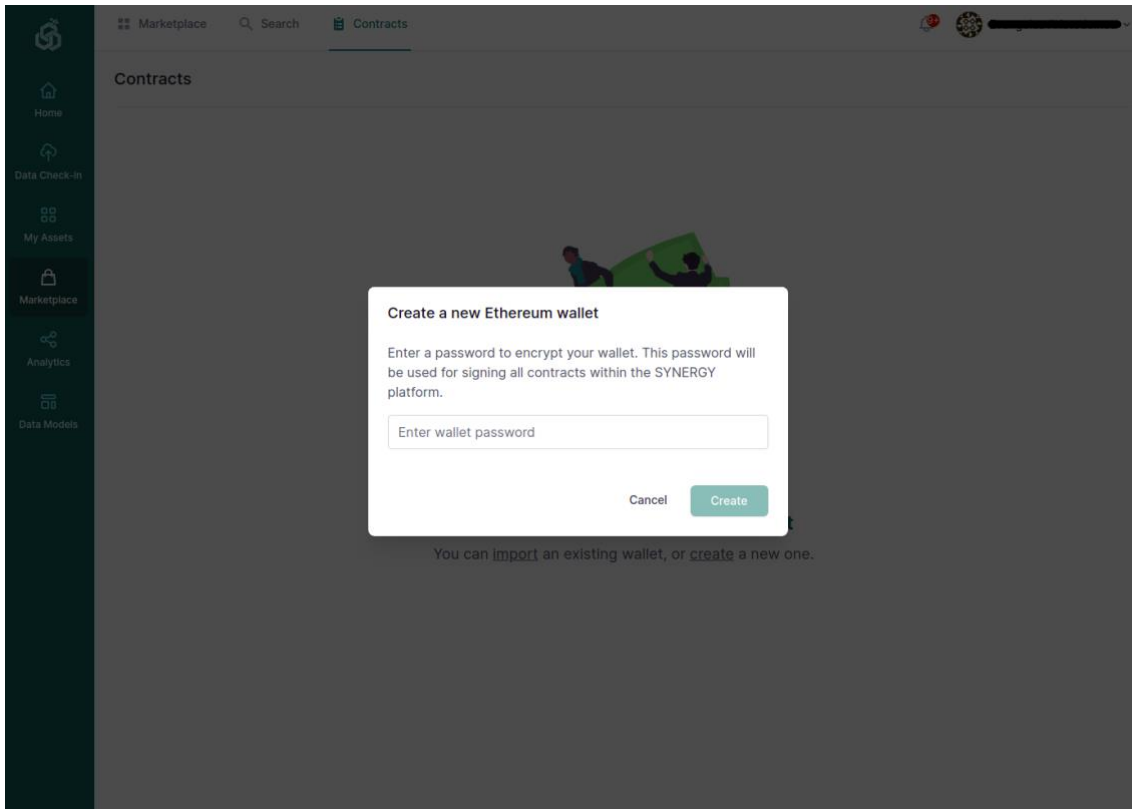


Figure 106: Create a New Wallet

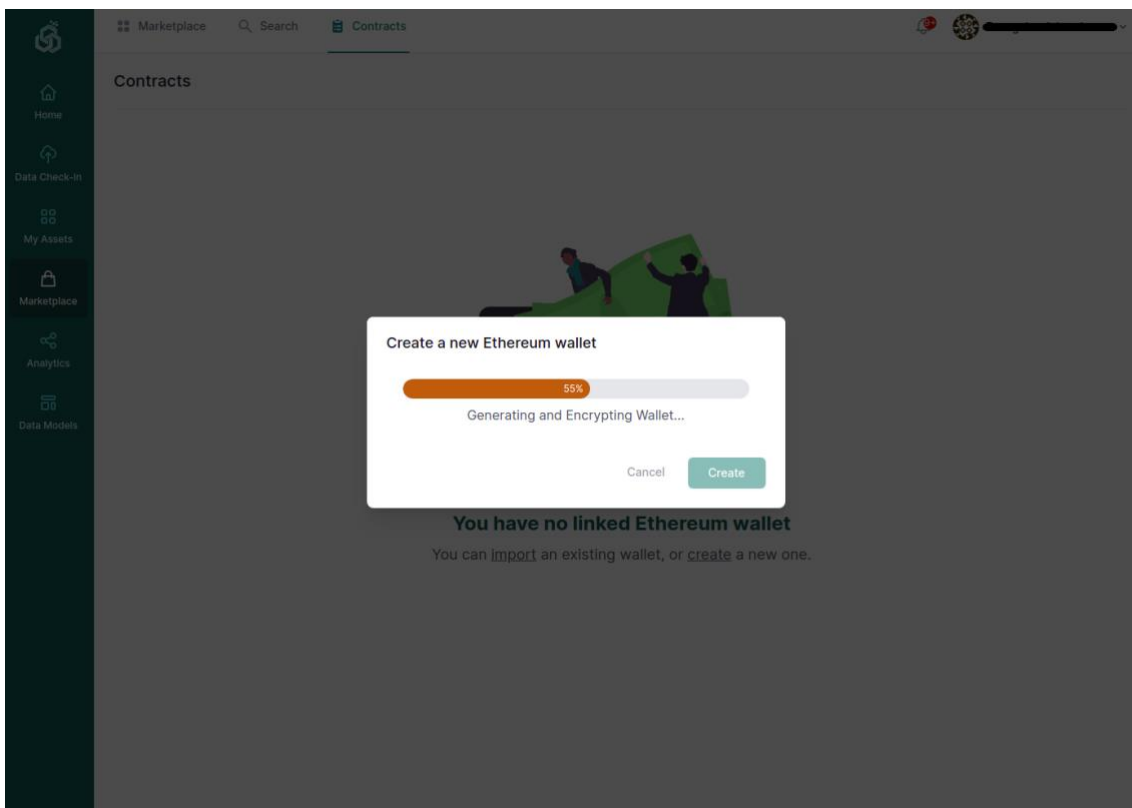


Figure 107: Create a New Wallet - Generating and Encrypting Wallet

6.4.3 View Wallet Details

At any moment, the organisation’s manager may view the wallet’s balance and the private key (in case it needs to be used in another application).

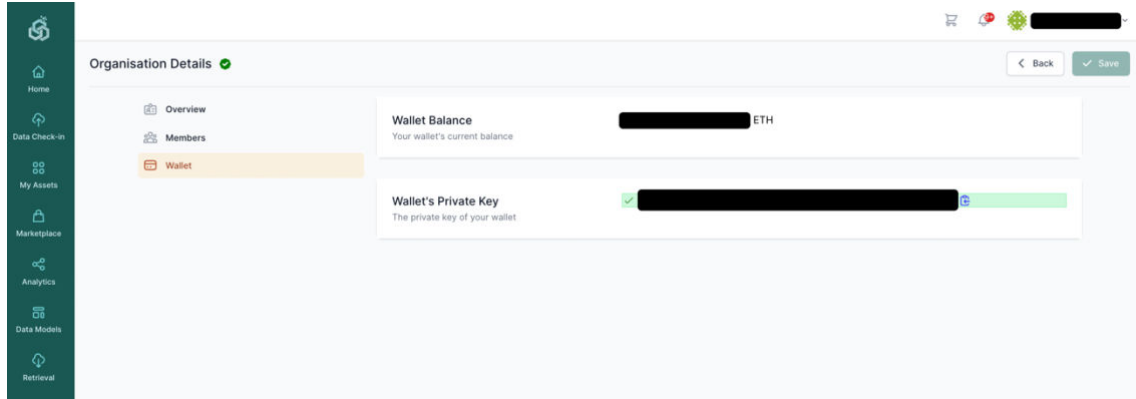


Figure 108: Wallet Details

7 Platform Integration and Support Activities

This section focuses on the integration approach followed by the technical partners to deliver the first official release of the SYNERGY Integrated Platform, as well as on the support channels that are in place to acquaint the overall SYNERGY consortium with the platform functionalities.

7.1 Integration Approach

As described in Section 2, the SYNERGY Platform brings together an extensive number of components, services and technologies that have been integrated at the Cloud deployment and the Server and Edge On-Premise Environments in accordance with the SYNERGY architecture (described in the SYNERGY Deliverable D2.7) and the integration plan (defined in the SYNERGY Deliverable D3.4). In order to proactively meet the software delivery challenges that a complex project as SYNERGY brings, an agile integration approach based on DevOps and GitOps has been put into place:

- I. **Release design and planning** that bring together the involved development teams, the WP3 and WP4 leaders, and the Technical Coordinator to discuss the priorities, the identification of dependencies and the main milestones per development cycle.
- II. **Collaborative software development and testing** with each development cycle (for new features) typically lasting 3 weeks and providing a new minor release of the SYNERGY Platform. It involves the involved development teams in: (a) coding, (b) running unit tests, (c) committing code to the respective component repo, (d) reviewing the code changes and providing comments to the pull requests, (e) merging the new features and fixes included in the code, (f) running appropriate integration tests.
- III. **Deployment in staging environment** that ensures that any minor release of the SYNERGY Platform is tested by the involved development teams before it is shipped to production. Depending on its criticality, any error or bug noticed may be prioritised to be immediately fixed or put in the backlog for the next minor release in (II).
- IV. **Deployment in production environment** that occurs per major release (i.e. on M24 for Release 1.00 and on M36 for Release 2.00) and on selected minor releases (for bug fixes or for prioritised features that need to be shipped earlier than the next major release). Experimentation on the production environment is available to all SYNERGY partners.



In this context, the integration activities of the SYNERGY Platform rely on a set of state-of-the-art techniques and open-source tools to ensure collaborative and continuous planning, development and deployment cycles for each release as depicted in the following table.

Table 3: SYNERGY Integration Tools

Integration-related Activity	Tool
Source code versioning and issue tracking	Github
Automated build and testing	Github Actions
Error Tracking	Sentry
Deployment	Docker, Kubernetes
GitOps	Flux

Taking into consideration the current status of the first official release of the SYNERGY Platform and the components interrelations, the integration plan has generally proceeded according to plan with minor deviations (i.e. the Remuneration Payments had to be brought earlier on the M24 release while the Support for the Wallet Manager in OPE and the integration of the Matchmaking Engine have been postponed for the M36 release) as depicted in the following figures.



Figure 109: SYNERGY Integration Plan on M24-Part I



Figure 110: SYNERGY Integration Plan on M24-Part II

7.2 Platform Availability

The SYNERGY Platform is deployed in its production environment at:

<https://www.synergy-bigdata.eu/>

Since the SYNERGY Platform allows organisation-based access, the organisation manager needs to register his/her personal profile and then register the organisation. Taking into consideration that the SYNERGY Platform is addressed to stakeholders of the electricity data value chain, the platform administrator moderates the organisation’s registration and may grant or deny access to the platform. If an organisation is approved, then the organisation’s manager can invite additional members to join the organisation, which receive an invitation code via email.

Demo accounts may become available upon request at: admin@synergy-bigdata.eu

7.3 Platform Support Mechanisms

In order to provide efficient support to the demo partners in WP8 and the energy application developers in WP5-WP7, the following support mechanisms shall be leveraged:

- **Online training sessions** aiming at providing walkthroughs and guidance of the SYNERGY Platform functionalities to the demo partners and the application developers. The first

platform demo event occurred on July 7th, 2021 in which the SYNERGY Beta Platform was launched to the whole SYNERGY Consortium. Various demos were provided in the bi-weekly technical calls while the SYNERGY Release 1.00 is to be formally introduced to the whole consortium in the SYNERGY Plenary Meeting on February 9th, 2022.

- **Direct support channel in Slack** under a “SYNERGY-Support” workspace in which dedicated channels have been created:
 - *#platform-app-support* intended for the energy apps developers to request support and guidance on the use of SYNERGY Platform by the technical team.
 - *#platform-demo-support* intended for the demo partners to request support and guidance on the use of SYNERGY Platform by the technical team.
 - *#platform-deployment* intended for the technical team to announce maintenance times and any redeployments that occur on the SYNERGY Platform, to the SYNERGY Consortium.
 - *#pretrained-analytics* intended for the technical partners’ interactions concerning the WP4 analytics activities.
- **Issue tracking in Github** under a dedicated organization entitled “SYNERGY Project”. A “*platform-support*” private repo has been created and includes pre-defined templates: (a) Bug report to collect all the bug information that will help the technical team to reproduce and solve the issue encountered (as depicted in the following figure); (b) Feature report to collect any enhancements and ideas that can improve the user experience and can be considered by the technical team for the next releases of the Platform. In order to properly manage, prioritise and track any problems/issues until they have been successfully resolved, all partners have been invited to join the SYNERGY Project in Github. It needs to be noted that issue reporting via email is not allowed.

It needs to be noted that additional support mechanisms are planned as follows: (a) Dedicated Github repo for external stakeholders that will obtain access to the SYNERGY Platform as part of the Living Labs activities, (b) Online documentation under a Help menu in the SYNERGY Platform in order to provide detailed guidance on the use of each functionality (for which a high-level introduction was provided in this deliverable).



8 SYNERGY Baseline Analytics

8.1 Baseline Data Analytics

The SYNERGY Deliverable D4.2 has already provided a detailed report on the scope and design of the SYNERGY Baseline Analytics, and explained the way the solutions being implemented in this context will gradually populate the SYNERGY analytics catalogue. As explained, the population of the catalogue with analytics solutions and the definition of the problems that they will target, will be continuous and performed in alignment with the use cases elaborated in T2.1, the demonstrators' data assets profiled in T2.3, the energy applications needs, as defined in WP5, WP6 and WP7, and the demo cases executed in WP8. Therefore, two additional intermediate releases have been planned for the baseline analytics, the first of which is provided in the current deliverable.

The deliverable at hand reports on the updates of the SYNERGY baseline analytics solutions that have been performed since the initial release documented in D4.2. It should be noted that the updates at this phase correspond to additional analytics solutions that fall under three main categories: (a) solutions to problems that had not been tackled in the previous release, (b) alternative solutions to previously explored problems that aim to enrich the available offerings, potentially addressing slightly different problem formulation and types of input data and (c) complementing services to solutions reported in D4.2. It is expected that improvements on the corresponding models and analytics workflows implemented until now will be made available in the next release of the analytics catalogue, which will be reported on M33, as then concrete insights from the actual available data will be integrated in the provided solutions and the analytics catalogue will be adapted to more concise needs of the SYNERGY stakeholders.

Although there is no change on the rationale behind the analytics portfolio and the way the catalogue is structured and its models are categorised, for completeness the two main categorisation axes of the SYNERGY baseline analytics solutions are again presented below:

Axis I: The analytics problems considered in this context fall under two broad categories depending on the scope of their application:

- **Personal Analytics** which refer to energy problems evolving around consumers (or small groups, such as a family) e.g., identification of consumer energy consumption patterns and preferences.



- **Industrial analytics** which refer to energy problems at a higher-level - at the building/ district/ grid level, where individual consumers and behaviours are not relevant e.g., energy demand forecasting at grid level and RES plant maintenance analytics.

Axis II: A second more fine-grained grouping of the analytics problems, this time focusing on the nature of the underlying problem has been devised as follows:

- I. **Demand Forecasting**, which refers to problems of energy demand forecasting at various levels and horizons.
- II. **Generation Forecasting**, which refers to problems of energy generation forecasting at various levels and horizons.
- III. **Occupants' Behaviour and Comfort Profiling**, which refers to problems around energy consumption patterns linked to consumers' thermal and visual preferences.
- IV. **Flexibility Forecasting**, which refers to the identification of the portion of demand/generation that can be reduced, increased, or shifted within a specific duration.
- V. **Predictive and Preventive Maintenance**, which refers to proactively identifying when maintenance on energy-related assets (RES plants, grid assets) should be performed.

As stressed in D4.2 the groupings along the two axes do not have one-to-one mapping and cannot be used in a hierarchical way. As an example, demand forecasting problems can be considered both in the context of personal and industrial baseline analytics, e.g., household demand forecasting vs grid-level demand forecasting, respectively.

D4.2 presented the analytics solutions that SYNERGY aims to provide, which were identified in the form of questions of interest for the electricity value chain stakeholders. The analytics solutions that are presented in the next sections, further broaden and complement the functionalities provided by the SYNERGY baseline analytics catalogue. Each of the solutions presented here refers to and addresses specific question(s) from the ones defined in D4.2 and this information is provided in the introductory table within each solution's sub-section.



8.2 Baseline Data Analytics Updates

8.2.1 AC consumption for different comfort levels

Scope	Personal Analytics
Problem	Occupants’ Behaviour and Comfort Profiling, Demand Forecasting
Question	<p>PERS.III.7 - What is the expected usage (in terms of energy consumption) of HVAC device for an occupant/ group of occupants in the next hour(s) to remain within specific comfort boundaries?</p> <p>PERS.III.11 - What is the expected energy consumption by a specific occupant/ group of occupants in the next hour to maintain the same comfort level?</p>

8.2.1.1 Description

Consumers’ engagement in Demand Response programs requires that their comfort level is maintained minimizing the risk of feeling uncomfortable during the demand response event. Equally important is the prior knowledge of their energy consumption during the event. The financial benefits by reducing the energy demand are evaluated together with the possibility of their comfort being compromised, so that an informative decision can be taken about participating according to consumers’ preferences. In recent years, a lot of research has focused on the estimation or the prediction of the AC energy consumption for different comfort levels, i.e., desired thermal conditions, since air conditioners are responsible for an average of 45% of residential electricity consumption (Lork et.al., 2017).

8.2.1.2 Background

Along the years, different approaches have been employed for the estimation of AC energy demand while in parallel taking into consideration the comfort of the consumers. One of the first attempts towards this direction and the most straight-forward is the formulation of a mathematical model based on the physical characteristics and the operation schedules of a building along with the weather conditions of the area. In (Wang et.al., 2013), the authors suggested a model, based on the building characteristics of a data centre, the weather conditions, and the desired room temperature (user comfort) to compute the energy demand of the AC system. The results of their research were promising, however the formulation of such equations requires a lot of information about the building under study, many assumptions need



to be made and the final model is too specific describing the energy needs of buildings with similar profiles. (Jain, M et.al., 2016) suggested a more general approach that did not take into consideration building-specific information. The energy consumption of the AC for a certain duration and desired temperature was calculated by the multiplication of a state vector (the states of the AC, on/off) with the nominal power of the AC. They suggested a model for the extraction of the state vector based on the indoor, outdoor and the desired temperatures. Even though this approach is more general, and the reported results are sufficient, there are some limitations concerning the sensors' position in the room, the closer to the AC the more accurate the extraction of the AC state vector. A more data-driven approach was suggested in (Lork et.al., 2017), where the authors tried to predict the consumption of the AC for the next 15 minutes. To that end, they trained different machine learning models after having engineered features related to the weather conditions, the desired and actual indoor temperatures, the state and the consumption of the AC. Another approach a little different from the aforementioned ones is the optimization method. (Khorram et.al, 2020) and (Papadopoulos et.al. 2019) formulated energy consumption equations and tried to minimize them considering at the same time temperatures constraints that represented the comfort of the users.

8.2.1.3 SYNERGY Implementation Details

8.2.1.3.1 *Input Data*

The data used for the proposed approach were acquired from the Indian Dataset for Ambient Water and Energy (iAWE) project. Thirty-three (33) sensors were placed for a period of 4 months (May-August 2013) in a three-storey house in New Delhi. The sensors collected electricity consumption data of different appliances, as well as water and ambient related data at a frequency of 1 second. After an initial processing of the dataset (data cleaning, missing data imputation), the resulted dataset spanned over a period of 3 months (June-August 2013) with a total size of 5,098,699 samples. Electricity consumption of ACs and ambient related data were used for the implementation of our approach.

8.2.1.3.2 *Approach*

The scope of our approach is to compute the AC energy consumption for a specific desired indoor temperature and duration of time. The idea is based on the implementation suggested in (Jain, M et.al., 2016). Two models were built based on features related to the AC electricity consumption and the indoor and outdoor ambient data. The first model (thermal



model) is a regressor that predicts the indoor temperature change for the next time interval and the second model is a classifier (state predictor) that predicts the next AC state (ON-OFF) based on the thermal model's output and the desired temperature.

Step 1: Feature pre-processing

- Resample the dataset (e.g., change the resolution from 1s to 2min)
- Extract and use the data from time intervals during which the AC is turned ON
- Select only the features related to the task and create new ones as needed
- Feature normalization

After the feature preprocessing phase, the number of samples in the dataset reduced to 42489.

Step 2: Model training, evaluation

For the training and the evaluation of the 2 models the dataset was split into training (80%) and test sets (20%). A Random Forest classifier and a Linear Regressor were trained for the implementation of the AC state predictor and the thermal model respectively. For the evaluation of the state predictor F1, recall and precision metrics were used, while MAE and MSE were used for the evaluation of the thermal model.

8.2.1.3.3 Technology

The 2 models were trained using the Sklearn library outside the SYNERGY platform, then saved as pickle files and registered in the platform. The version of the library used is the same as the one provided by the SYNERGY platform. After the model registration, a pipeline was built in the SYNERGY platform, where the trained models were applied in an appropriate way for the needs of the task.



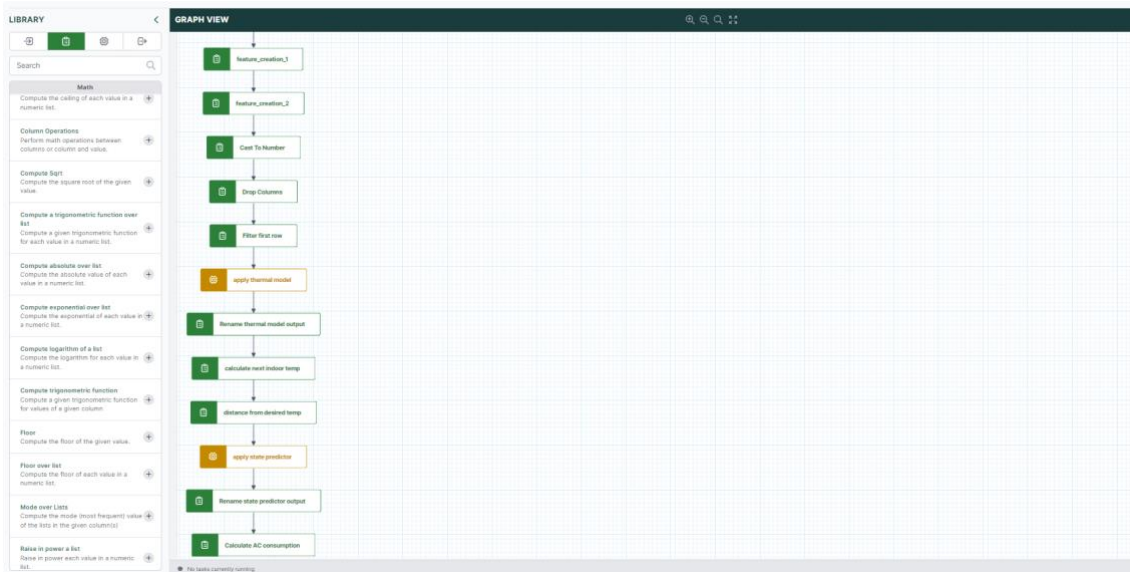


Figure 111: Model Application Pipeline (partial view)

8.2.1.3.4 License & Availability in the SYNERGY Platform

The trained models are available in the SYNERGY Platform, under the names “thermal_model” and “State_predictor”. They can be accessed through the SYNERGY marketplace and used in an analytics workflow in a similar way to the one presented in Figure 111, under the terms of the established contract between the data asset provider (Suite5) and consumer.

The aforementioned workflow is also accessible in the SYNERGY platform by the name “AC consumption at different comfort levels (apply)” to facilitate users who obtain the aforementioned models. The trained models and the pipeline are provided with a confidential license by their provider, Suite5, that is the owner of the models’ intellectual property rights.

8.2.1.3.5 Assumptions and Limitations

The task presented at this section aims to predict the AC energy consumption when regulated to satisfy specific comfort requirements (i.e., maintain the indoor temperature to certain degrees). The proposed approach exploits information coming from the small fluctuations in indoor temperature, which can best be captured on datasets gathered using short sampling periods (seconds to 5 minutes intervals).

The features that were selected for the training of the models relate to the AC energy consumption as well as to indoor and outdoor environmental data. New features can be added in the future, such as datetime and other environmental data in order to assess if they can improve the performance of the overall approach.



8.2.2 AC consumption flexibility forecasting

Scope	Personal Analytics
Problem	Occupants Behaviour and Comfort Profiling, Flexibility Forecasting
Question	PERS.III.13 - What is the expected (very) short-term demand flexibility at device level within comfort boundaries?

8.2.2.1 Description

Balancing the electricity grid in a cost-efficient way, avoiding heavy investments in new power plants and transmission lines is substantial, as the energy demand in recent years increases. Residential buildings are responsible for a significant amount of this increase (Alic, O., & Filik, Ü.B., 2020) and especially air conditioners are responsible for an average of 45% of domestic electricity consumption (Lork et.al., 2017). Demand response programs aim to achieve grid balancing by reducing the electricity demand during peak hours of a day. As mentioned in section 8.2.1 however, participating in Demand Response programs should not compromise residents' comfort. Thus, estimating consumers' flexibility, i.e., the modification of the energy consumption in reaction to an external signal without risking discomfort (Eurelectric, 2014) is essential.

8.2.2.2 Background

Most of the research focused on consumers' flexibility quantification require the formulation of one or more theoretical models that can describe the flexibility quantification of appliances, the thermal characteristics, and the energy needs of the building under study. These models can be used to estimate the electricity flexibility for a certain period by calculating the difference of the electricity demand for different input settings (desired indoor temperature). Chen, Y., et.al., (2019) formulated models to estimate the potential flexibility that different resources of a building can offer (e.g., thermal mass, appliances, HVAC system, water tank). Alic, O., & Filik, Ü.B. (2020) formulated a dynamic model that expresses the correlation among the power consumption, temperature, and time of ACs activations. They also consider different pricing schemes based on which different AC settings (temperatures) were chosen. During high pricing periods lower temperatures were selected to reduce the AC consumption while keeping the consumers comfort among acceptable ranges. The authors of (Che, Y., et.al., 2019) proposed an



electric model to represent the operation of an AC and suggested an approach to maintain the indoor temperature stable without compromising resident’s comfort, while reducing the energy demand.

8.2.2.3 SYNERGY Implementation Details

8.2.2.3.1 *Input Data*

The Indian Dataset for Ambient Water and Energy (iAWE) was selected for this task, same as in section 8.2.1, as the solution in this problem leverages the solution implemented to predict AC consumption at different comfort levels, i.e. different thermal conditions, which was presented there.

8.2.2.3.2 *Approach*

To unlock consumers’ flexibility from air conditioners, this approach employs the trained models of section 8.2.1 and the occupants’ thermal comfort boundaries estimated by the approach described in section 3.3 of D4.2. The already trained models of section 8.2.1 can be used to estimate the minimum and maximum AC energy consumption corresponding to the comfort boundaries, from which a specific user’s flexibility can be obtained.

Step 1: Feature-preprocessing

Perform the required steps as explained in section 8.2.1.3.2

Step 2: Extract the comfort boundaries

At this step the user’s comfort boundaries are defined based on the approach described in section 3.3 of D4.2. The consumer’s comfort profile is extracted as the minimum and maximum acceptable temperatures. Depending on whether the AC is used for heating or cooling, the minimum or the maximum temperature respectively will be used to compute the flexibility.

Step 3: Flexibility estimation

As a final step, the trained models of section 8.2.1 are employed to estimate the AC energy consumption for both the optimal temperature and the temperature that is farthest from the optimal but still acceptable (minimum/maximum), as these are defined by the user’s comfort profile and for a selected time duration. The difference of these values equals the flexibility that can be offered to the market during this period.



8.2.2.3.3 Technology

All the models and the dataset used for this task, as mentioned before, are uploaded and available on the SYNERGY platform, following the guidelines for the versions of the required libraries. Next, a pipeline was created on the platform, where a short data pre-processing is performed, and the models are applied. The libraries needed for the pipeline implementation are Sklearn and Pandas and their versions are the ones provided by the platform.



Figure 112: Analytics Application Workflow (partial view)

8.2.2.3.4 License & Availability in the SYNERGY Platform

As explained, the current implementation does not implement and train any new machine learning models but constitutes an application and appropriate combination of other solutions that are available in the analytics catalogue. A workflow has been created that implements this application logic and provides the final result, i.e. the predicted flexibility. Part of the workflow is presented in Figure 112 and is indicative of the required pre-processing of the input data. The workflow is accessible and configurable in the SYNERGY platform by the name “AC consumption flexibility forecasting”. The created workflow is provided with a confidential license by its provider, Suite5, that is the IPR owner of the specific workflow but also of the leveraged models.

8.2.2.3.5 Assumptions and Limitations

The presented approach employs already trained models and assumes that they can adequately estimate the comfort boundaries and the AC consumption based on the input data. The limitations of the individual models have been presented in the corresponding sections.

However, in future application scenarios the input data may deviate significantly from those used to train the models, rendering model retraining a necessity.

It should also be noted that the thermal model presented in this approach is data-driven and does not take into consideration parameters related to building characteristics. Instead of the proposed model, other types can also be used, either data-driven based on a different set of input features or building-based thermal models, that describe how temperature changes in relation to building characteristics.

8.2.3 Anomaly Detection in household energy consumption

Scope	Personal Analytics
Problem	Occupants Behaviour and Comfort Profiling, Demand Forecasting
Question	PERS.III.9 - Are the occupant’s actuations in accordance with the expected energy consumption patterns and comfort preferences?

8.2.3.1 Description

Global warming and the continuous rise in worldwide demand for energy have made energy efficiency a top priority for many countries. With most of the total energy still being produced by fossil fuels, reducing the total energy consumption percentage is very crucial along with finding more sustainable energy sources. While buildings, in general, are responsible for a huge amount of energy consumption (Gul et al, 2015), a study showed that home electricity consumption accounted for 38.9% (1.46 trillion kWh) of the total annual electricity consumption in the United States (EIA, 2022), making the residential sector the world's primary energy saving target among end-use sectors.

Inefficient management and unawareness play an important role in wasting energy resources, so promoting energy-saving procedures and appropriate usage of household appliances is of utmost importance. One promising approach to improve energy efficiency is to identify anomalies in building energy consumption. With the implementation of energy monitoring systems, anomaly detection techniques can help detect possible abnormal consumption patterns, faulty equipment, power theft and aid in better decision-making.

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior (Chandola et al., 2009). Nowadays, the amount of data generated by sub-



meters and smart sensors makes the problem of detecting anomalies and localizing them very challenging. If we know the day-to-day energy usage patterns, it is possible to analyze sudden and unexpected changes by processing these data and optimizing energy consumption by providing feedback to the occupants. Moreover, since electricity bill prices are constantly increasing, by detecting and preventing wasting energy, end-users' affordability is promoted.

However, there is no obvious definition of normal and anomalous energy consumption and the difference between normal and anomalous behaviors cannot always be determined by some unified metric. That, in addition to the lack of ground-truth data, makes the various anomaly detection techniques very difficult to evaluate.

8.2.3.2 Background

Recent research in the past few years suggests that a significant effort has been put into innovating anomaly detection techniques for time-series data (Braei et al., 2020) and specifically on the energy field (Himeur et al., 2020).

First of all, three types of anomalies exist: point anomalies where a point significantly deviates from the rest of the data; collective anomalies where an individual point is not an anomaly, but a sequence of points can be labeled as such; contextual anomalies where points are considered as anomalies in a certain context and as normal in another.

Furthermore, a categorization of the anomaly detection methods can be made taking into account if the data are labeled or not: supervised anomaly detection considers each timestamp to have a label indicating whether it is an anomaly or not; semi-supervised methods can be used when only normal time-series exist for training; unsupervised techniques assume that the dataset is unlabeled. However, the lack of annotated datasets for anomaly detection has given unsupervised methods a more practical use in real-world operational data despite the main limitations of these methods like: i) performance & computational efficiency when scaling to big data; ii) performance reliance on feature selection (Fan et al., 2018).

Another classification can be made based on the methods employed to detect the anomalies: statistical methods; classical machine learning methods; deep learning methods.

Statistical methods usually include the Moving Average model (MA) where the residuals after the model fit are used to determine an anomaly score. In a similar manner, other statistical models are utilized like the Autoregressive (AR) and the Autoregressive Moving Average model (ARMA) although they consider some assumptions about the stationarity of the data. A recent



example is a variation of Seasonal Trend decomposition using Loess (STL) which produced promising results on a real-world dataset (Lee et al., 2017). Popular machine learning techniques include various clustering algorithms in order to classify power consumption data into normal or abnormal. A novel study proposed the k-means algorithm to separate anomalies from normal events in large energy log files (Henriques et al., 2020). A one-class support vector machine (OCSVM) was used for the detection of anomalous events or actions in a real-world smart home dataset (Jakkula et al., 2011). Deep learning methods, like the autoencoder, have been receiving more attention lately. An important property of autoencoders is the fact that they can employ nonlinear dimensionality reduction, as was practically demonstrated in a recent study using artificial data generated from a complex nonlinear system (Sakurada et al., 2014). In another study, a deep autoencoder architecture was used in a semi-supervised problem to detect abnormal patterns in building energy data. It was first trained on a normal dataset and then used to determine anomalies using reconstruction errors. Additionally, it was shown that a small proportion of masking noises can enable autoencoders to learn more reliable and robust features from real-world data (Fan et al., 2018). An Autoencoder was also used to detect abnormal days in building energy consumption data (Chahla et al., 2019) and then a combination of LSTM and K-means algorithms was used for the localization throughout the day. In a similar vein, interesting research proposed the combination of two anomaly detection frameworks, a pattern-based anomaly classifier based on the autoencoder and an ensemble of prediction-based anomaly classifiers based on support vector regression and random forest. The evaluation was performed using real-world energy usage data generated from buildings and showed that the ensemble framework significantly improved the performance (Araya et al., 2017).

An important consideration for anomaly detection in household energy consumption is that abnormalities typically are difficult to be detected when they are due to seasonal changes or other personal settings (holidays, family events). Another limitation is related to the aggregation level of the dataset. With aggregated consumption data, anomalies are hard to be detected, therefore appliance-level data are more appropriate to provide accurate information about the causes of each anomaly.



8.2.3.3 SYNERGY Implementation Details

8.2.3.3.1 *Input Data*

The dataset used to train and evaluate the model for anomaly detection was taken from The Building Data Genome 2 (BDG2) (Miller et al., 2017), an open data set made up of 3,053 energy meters from 1,636 buildings. The time range of the times-series data is two full years, the frequency is hourly, and the dataset includes measurements of electricity, heating and cooling water, steam, and irrigation meters. For our specific use case, we only used data gathered from a specific household building and only for electricity. The data were pre-processed, we imputed missing values by the average grouped by year-month-weekday-hour for the specific timestamp. The above filters resulted in obtaining a dataset of 17544 records, with the following 2 variables:

- Observed datetime, containing the timestamp of the specific measurement
- Energy consumption, containing the total energy consumption of the building (sum of KW per hour)

8.2.3.3.2 *Approach*

An anomaly detection method was implemented into a pipeline, with the aim to predict if the data for a specific timestamp represents an anomaly point.

Step 1: Feature preprocessing

Three new columns were created by extracting:

- the month of the year
- the day of the week
- and the hour of the day from the datetime column.

Step 2: Model training, evaluation

Using the energy consumption data and the new datetime features that we created, we utilized the One-Class SVM algorithm, that learns a decision function for anomaly detection, classifying new data as similar or different to the training set.

After we trained and applied the model to our dataset, the resulting column contained -1 for anomalies and 1 for normal points. We then mapped those values into a boolean column, indicating the anomalies.



8.2.3.3.3 Technology

The anomaly detection pipeline was created within the SYNERGY platform and therefore the platform’s libraries were used. The exact blocks that were used are shown in the following screenshot.

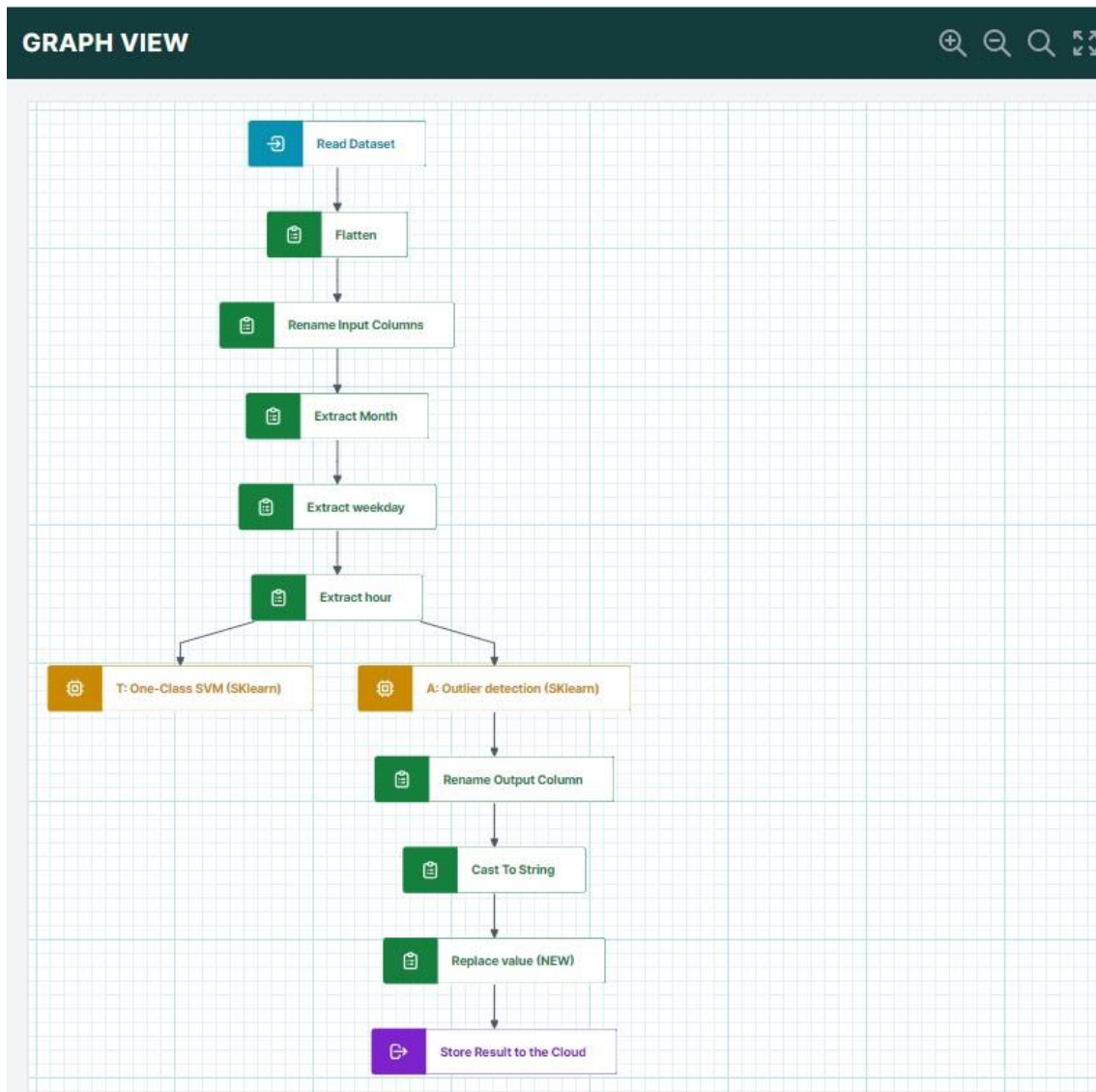


Figure 113: Model training and application (indicative) workflow

8.2.3.3.4 License & Availability in the SYNERGY Platform

The trained model is available in the SYNERGY Platform, under the name “Household_Energy_Anomaly_Detector”. It can be acquired through the SYNERGY marketplace and used in any pipeline through the “apply sklearn model for outlier detection” block. The pipeline presented in Figure 113 is available in the SYNERGY Platform under the name

“Household Anomaly Detection”. The trained model and the pipelines are provided with a confidential license by its provider, MAG, that is the owner of the model’s intellectual property rights.

8.2.3.3.5 Assumptions and Limitations

This model can accurately be applied for anomaly detection in new data generated from the same building, or a very similar one in terms of energy consumption behaviour.

New features can be added, like the presence of a national holiday or weather features and their contribution will be evaluated in order to improve the model performance in future releases. The trained model is available to be used within the SYNERGY platform and the description of the training dataset can help interested stakeholders understand whether their intended usage will be a good fit for the already trained model.

8.2.4 Building Energy Demand Forecasting

Scope	Industrial Analytics
Problem	Demand Forecasting
Question	IND.I.7 - What is the expected (very) short-term energy demand at building level?

8.2.4.1 Description

The building sector produced nearly 40% of total CO₂ emissions and consumed nearly a third of the global energy in 2018 (IEA, 2021). Moreover, rapidly growing urbanization, depletion of natural resources, and the continual increase of world population are some of the main reasons that improving energy efficiency in buildings is of vital importance. Energy demand forecasting could be used to achieve energy saving, optimize energy distribution plans and select proper energy conservation methods.

8.2.4.2 Background

As the application of the data-driven methods on building energy demand forecasting draws more attention, a variety of in-depth review articles have been published over the last few years (Sun et al., 2020; Bourdau et al., 2019). Studying these articles shows us that there is an agreement between all authors that, according to the forecasting accuracy measures, none of



the methods outperforms the others in all situations. Accordingly, data-driven approaches can be categorized as statistical models and machine-learning methods.

In the case of statistical models, further classification can be made into statistical regressions & time-series-based models. Statistical regressions usually include multiple linear regressions (MLR), elastic net regressions, Gaussian process regressions, etc. (Fan et al., 2017, Rahman et al., 2017) Moreover, common time-series-based techniques involve auto-regressive moving average (ARMA) and auto-regressive integrated moving average (ARIMA) models, but these only consider recent historical load demand data points to predict the future. To overcome this issue, exogenous variables have been included (ARIMAX) like weather conditions or occupancy profiles. Additional seasonal trends can be incorporated in the previous models to account for events that happen at a periodic pace (SARIMA, SARIMAX) (Chou et al., 2016; Cai et al., 2019).

On the other hand, the most prominent machine learning methods include the regression tree (RT), support vector regression (SVR), artificial neural network (ANN), deep neural network (DNN), and ensemble methods (bagging, boosting, etc.) (Fan et al., 2017). In a recent example, (Ding et al., 2017) utilized a variation of an SVR model together with historical and weather data to forecast one hour and one day ahead with 30 min intervals. Furthermore, a gradient boosting machine (XGBoost) was tested on a large dataset of 410 commercial buildings. The results showed that this model, when properly tuned using a k-fold blocks CV procedure, outperformed the industry's best practice model in most of the cases (Touzani et al., 2018). Further examination of gradient boosting methods for building energy demand forecasting has been made by (Bassi et al., 2021). Using a synthesized dataset, they demonstrated that XGBoost performed best compared to LightGBM and CatBoost.

8.2.4.3 SYNERGY Implementation Details

8.2.4.3.1 *Input Data*

The dataset used in the implementation was taken from The Building Data Genome 2 (BDG2) (Miller et al., 2017). For the current implementation, we only used data electricity consumption data gathered from a specific residential building. The following 3 variables were used:

- Datetime, containing the timestamp of the specific measurement
- Hourly consumption, containing the total electricity consumption of the building (sum of KW per hour)



- Temperature, containing the air temperature on the building's location for the specific timestamp in degrees Celsius (°C)

8.2.4.3.2 Approach

The goal of this analysis is to forecast the electricity consumption one hour ahead using historical data, temporal features, and the next hour's forecasted air temperature. The implementation was made using three separate pipelines, one for preprocessing, one for training and evaluation on the training set, and the last for applying the trained model and evaluating on the test set.

Step 1: Feature preprocessing

The following preprocessing steps were performed in the SYNERGY platform, to prepare the dataset for training a one-step forecasting model:

- Creation of new temporal features (month, weekday, hour)
- Creation of 24-hour lags of the energy consumption to be used as historical input values

Step 2: Model training

To capture the non-linear relationship between the available input parameters and the forecasted values, we utilized a popular ensemble method from the literature, the Gradient Boosting Machine (GBM). This is a very flexible, fast and robust method especially when working with high-dimensional input parameters. In addition, it allows the inclusion of non-influent variables without degrading performance. Specifically, we trained an extreme gradient boosting (XGBoost) model.

In our experimental setup, we split the data into a training and a test set in a 0.7-0.3 ratio. We first trained the model on the training set, and then used this model to make predictions on the test set. The resulting column contained predictions for the next hour's energy consumption.

Step 3: Evaluation

In conclusion, we evaluated our results using the mean absolute error (MAE), mean squared error (MSE), and explained variance, all popular evaluation metrics from the related literature. The output that we get is next-hour predictions of electricity consumption.





Figure 114: Indicative visualisation of forecasted vs actual values

An example is presented in Figure 114, where the predicted output of electricity consumption (in KW per hour) versus the actual consumption is depicted.

8.2.4.3.3 Technology

The pipelines for building energy demand forecasting were created within the SYNERGY platform and therefore the platform’s libraries were used. In the following figures, the blocks that were used for the three pipelines are shown.

- Preprocessing: Workflow for preprocessing and saving the datasets into the cloud

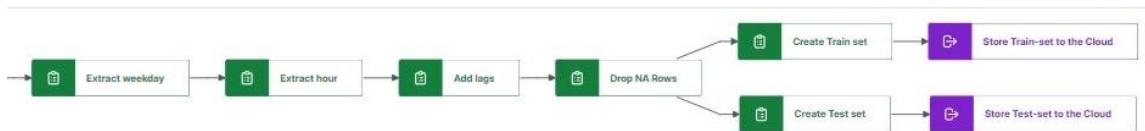


Figure 115: Preprocessing workflow (partial view)

- Training: Workflow for training the model using the training set



Figure 116: Model training workflow

_ Model Application: Workflow for applying the trained model to the test set & evaluating



Figure 117: Model application and evaluation workflow

8.2.4.3.4 License & Availability in the SYNERGY Platform

The trained model is available in the SYNERGY Platform, under the name “Building_Demand_Forecasting_XGBoost_Regressor”. It can be acquired through the SYNERGY marketplace and used in any pipeline through the “A: Regression labeled data” and “A: Regression on unlabeled data” blocks.

The pipelines presented in the previous section are available in the SYNERGY Platform and can be configured to be used with any appropriate dataset, offering guidance on how the model can be retrained on new data, or simply how data can be made usable by the model, either for training or application purposes.

The trained model and the pipelines are provided with a confidential license by their provider, UBITECH, that is the owner of the respective intellectual property rights.

8.2.4.3.5 Assumptions and Limitations

This model can be applied for predicting the next hour’s electricity consumption in new data generated from the same building, or a very similar one in terms of energy consumption behaviour. Also, this method can be used with data of different granularity, e.g. 15-minute steps, if retraining of the model is performed. However, availability of historical data for training the model is crucial since the input variable that contains the consumption of the previous time-step (lag-1) is the strongest predictor for next-step forecasting.



New features can be added, e.g. occupancy presence, and their contribution will be evaluated to improve the model’s performance in future releases.

The trained model is available to be used within the SYNERGY platform and the description of the training dataset can help interested stakeholders understand whether their intended usage will be a good fit for the already trained model.

8.2.5 Prediction of day-ahead demand flexibility at building level

Scope	Industrial Analytics
Problem	Flexibility Forecasting
Question	IND.IV.1 - What is the expected (very) short-term demand flexibility at building level?

This section presents a solution on demand flexibility prediction problems at building level, that builds upon a demand prediction model developed under T4.2 and reported in section 3.18 of D4.2 (“Prediction of day-ahead demand flexibility at building level”). Specifically, the solution presented here, builds upon the aforementioned, previously developed ANN-based model (D4.2), that provided a day-ahead demand prediction at building level, to ultimately provide the short term (day-ahead) demand flexibility prediction at building level.

8.2.5.1 Description

Electrical load forecasting is a process used to predict the power demand of a specific system or subsystem and is primarily used for planning and scheduling processes of power/energy systems. Load forecasting applies to various applications ranging from O&M on a consuming energy device level (e.g. HVAC systems) through to building energy performance optimization and grid-level energy management. Depending on the time horizon of planning strategies, load forecasting can be divided into three categories: short-term load forecasting (1 hour to 1 week ahead), medium-term forecasting (1 week to 1 year ahead) and long-term forecasting (longer than a year ahead of the time of demand). More recent studies have introduced a fourth category, the very short-term forecasting used for load forecasting from seconds up to one day ahead (Hong and Fan, 2016). The particular load forecasting model utilized in this flexibility prediction solution addresses the short-term and specifically day-ahead horizon and is tailored towards the prediction of total demand in buildings.



This model is then leveraged to provide flexibility prediction solution based on a continuous forecasting routine which updates the output of the forecast utilizing Artificial Neural Network (ANN) technology and particularly a Long-Short Term Memory Model (LSTM).

8.2.5.2 Background

Regarding load (demand) forecasting models per se, the relevant background literature review has been reported in the respective section (3.18) of D4.2. In this section the technology background focuses particularly on demand flexibility models.

Forecasting flexibility in electricity demand on a building level, literature is limited to studies covering this topic from a very specific perspective. (Gorria et al., 2013) presented a mathematical model for forecasting the aggregated electricity demand of a group of domestic consumers signed up to an incentive-based demand management program, under which consumers receive signals that offer financial incentives for limiting their volume of consumption at time intervals when system peak demand is forecast. The resulting optimization model is a mixed-integer linear programming problem implemented in JAVA and is applied to a case study in which the objective is to limit consumption by a population of 15932 consumers from 15:00 to 17:45 on a specific summer day.

In 2020, (Finck et al., 2020) performed a similar study based on economic model predictive control strategy which has gained attention in building energy management systems. This case study demonstrates the application of an economic model predictive controller under real-time pricing, including day-ahead prices and imbalance prices. For real-time prices in balancing and spot markets, a method that presents a flexibility service to provide demand flexibility for a specific time window in advance has been introduced and tested under real-life conditions, which also included the stochastic behavior of occupants and the dynamic behavior of the building and heating system. During the test periods, the controller managed the total operational costs of the heat pump's electricity consumption and achieved a prediction performance of Root Mean Square Error between 0.17 and 0.22 kWh.

Flexibility, in terms of the amount of load which can be switched, could be approached by various potential ways, depending on the load composition, the controllability of each load device within the specific load pool, as well as any technical device constraints, commercial arrangements and user requirements. Here, we present a generic flexibility calculation model which may be used independently of any specific load characteristics and features, that aims to



define the available flexibility of a total installation's load, compared to a declared availability at a previous point in time. In effect, this provides the opportunity for the overarching decision-making system (be that a local energy management system, an electricity supplier, etc), to consider amends to pre-submitted declarations aiming at achieving particular goals within changing market signals.

8.2.5.3 SYNERGY Implementation Details

8.2.5.3.1 *Input Data*

The dataset utilised to train and evaluate the demand prediction model that is embedded in the demand flexibility prediction model comprises of:

- Hourly day-ahead Numerical Weather Prediction (NWP) data (uvi, temperature) for a building location in Athens, Greece. The NWP data were retrieved from OpenWeatherMap¹ utilizing the respective API offered by the particular web-service ("Open Call API") which exposes a number of weather metrics.
- Weekday (0-6): It was identified during the training development that the day-of-week information, contributes to the improvement of the model's accuracy as it reflects on the demand habits of the consumer throughout the week.
- Time of day (1-24): Similarly, it was identified during the training development that the time-of-day (hour) information, contributes to the improvement of the model's accuracy as it reflects on the demand habits of the consumer throughout the day.
- Hourly demand historical data of the building

8.2.5.4 Approach

Step 1: Feature pre-processing

The model includes the following pre-processing steps that apply on the dataset:

- Data normalization (scaling). A Standard scaler was utilized

¹ <https://openweathermap.org/>



- Dataset reshaping in an appropriate 3D matrix which is a hard requirement in LSTM models.

Step 2: Model training

The short-term demand prediction model that lays at the core of this flexibility prediction solution, falls under the category of long-short term recurrent neural networks (LSTM RNN). The Keras library was utilized to define the sequential class of the model comprising of an LSTM layer with a number of hidden cells followed by a Dense output layer. The model uses batches of 24 input-output pairs. The LSTM and Dense layers utilize different activation functions in order to address the specificities of the model's features.

Step 3: Model performance evaluation

The performance of the demand flexibility prediction model coincides with the performance of the embedded demand prediction model. The initially available dataset was split such that 80% was used for training and the rest 20% for the evaluation of the model. The training pipeline fed the model with 24-hour ahead input batches of weather forecasts, namely UVI and temperature as well as "day-of-week" and "time-of-day" information - unseen by the model during the training stage - which in turn return the predicted building demand. The model is compiled using an "Adam" optimizer which outperformed previous trials with stochastic gradient descent (SGD) optimizer. During the development phase a mean square error (MSE) loss function and "Accuracy" metrics were utilized to evaluate the model's accuracy when applying certain modifications such as the number of hidden cells, optimizers selection, activation functions etc.

8.2.5.5 Technology

The day-ahead demand prediction model is a Python model, developed and trained outside the SYNERGY environment. The basic libraries utilized for the training of the model are: Tensorflow-Keras, Scikit-learn, Pandas and Numpy. The versions of the libraries used for the process training are aligned to the model registration guidelines. The trained model was saved in "h5" format and registered in the SYNERGY Platform.

After registering the aforementioned model in the SYNERGY Platform, a pipeline was created through which the aforementioned model is applied in the context of a more complex workflow -as presented in the following figure- that enables the prediction of day-ahead demand flexibility at building level, through the comparison of an early demand prediction with an updated one (closer to realization).



The application pipeline comprises two, initially independent flows, each one utilizing different input datasets but both primarily resulting in the day-ahead demand prediction of the building. The critical difference between the two flows is that the one is based on early day-ahead weather forecasted input parameters, retrieved from an external weather service provider at 09:00 am of previous day (D-1), while the other is based on an updated day-ahead weather forecast, retrieved at 21:00 pm of D-1, from the same source. Apart from the predicted weather parameters, the input datasets feeding each flow and ultimately the demand flexibility model, also include “day-of-week” and “hour-of-day” features. Prior to the input datasets being ready to feed the demand prediction models they enter a pre-processing stage including normalization and other vetting processes in order to comply with the demand prediction models’ requirements. Ultimately, the demand flexibility model’s output is the signed flexibility in kW (upwards, downwards), calculated as a result from the comparison of the early demand prediction versus the updated one (closer to realization).

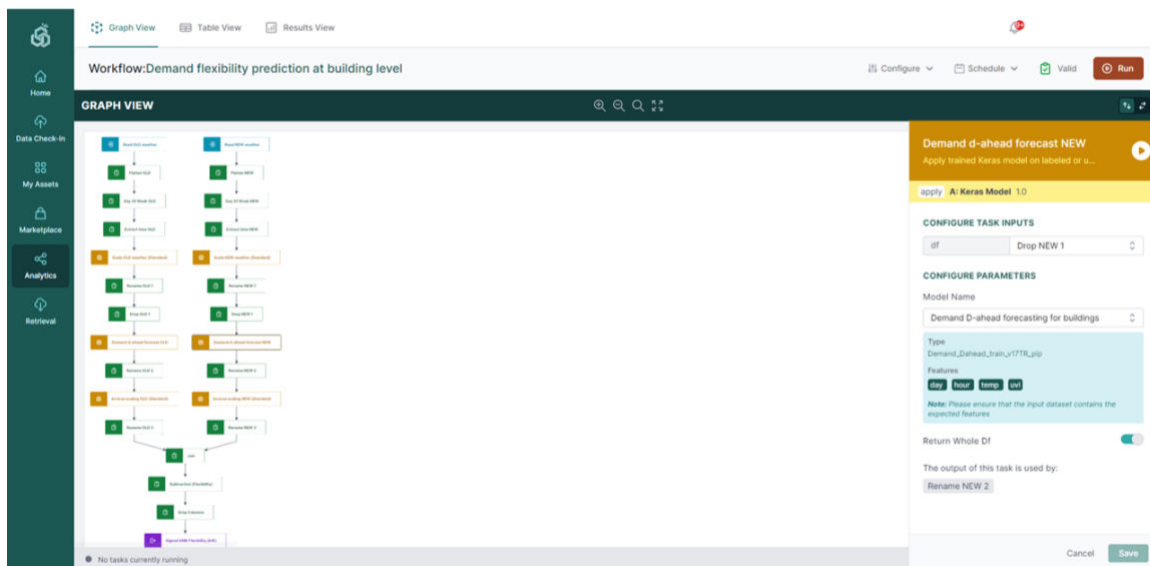


Figure 118: Model application workflow

8.2.5.6 License & Availability in the SYNERGY Platform

The trained model is available in the SYNERGY Platform, under the name “Demand D-Ahead forecasting for buildings”. It can be accessed through the SYNERGY marketplace and used in an analytics workflow in a similar way to the one presented in Error! Reference source not found., under the terms of the established contract between the data asset provider (VERD) and consumer.

The aforementioned workflow is also accessible in the SYNERGY platform by the name “Demand flexibility prediction at building level” to facilitate users who obtain the aforementioned model. The trained model and the pipeline are provided with a confidential license by their provider, VERD, that is the owner of the model’s intellectual property rights.

8.2.5.7 Assumptions and Limitations

The demand prediction model that lays at the core of the demand flexibility prediction model under discussion, is found to provide satisfactory results when used for residential buildings of similar climatic conditions and demand behaviour, its accuracy is expected to be limited when used for buildings of other types of use (e.g., commercial) or located in a wider geographical area, due to the localized nature of the training dataset. The description of the training dataset will be available through the platform, which can help interested stakeholders understand whether their intended usage will be a good fit for the already trained model. At a later stage of the project the possibility of making the training pipeline available in SYNERGY Platform will be investigated with the aim to enhance model’s accuracy by allowing it to be trained on users’ dataset.

8.2.6 Prediction of generation flexibility at DER level - short-term

Scope	Industrial Analytics
Problem	Flexibility Forecasting
Question	IND.IV.4 - What is the expected (very) short term generation flexibility at DER level?

This section presents a solution on generation flexibility prediction problems at DER level, that builds upon a PV generation prediction model developed under T4.2 and reported in section 3.20 of D4.2 (“Prediction generation flexibility at DER level”). Specifically, the solution presented here, builds upon the aforementioned, previously developed ANN-based model that provided a day-ahead generation prediction for PV installations, to ultimately provide the short term (day-ahead) prediction of generation flexibility at DER level.



8.2.6.1 Description

Solar PV forecasting techniques are primarily driven by the dynamic nature of solar irradiance and other relevant meteorological parameters, that induce high levels of uncertainty in the process of effectively harnessing solar energy. This dynamic nature is translated into voltage and power fluctuations with subsequent impacts on energy management microgrids, local distribution networks or even at a national grid level. From another point of view, accurate PV generation forecasting can be very important when predicting future availability levels of flexible resources such as batteries.

Standalone PV systems may offer two types of flexibility regarding their operation. On one hand, the available interfaces exposed through their power electronics systems, may offer the possibility of curtailing production by setting power export limits on the systems' inverters, or indeed by communicating specific setpoints for their operation which may be obtained by an external request and realized in combination with a PV system's maximum power point tracking (MPPT) capabilities. On the other hand, PV production forecasts may be used for market related actions (e.g. obligatory declarations of availability of a system, utilization in energy management and resource scheduling systems, etc), where continually updated forecasts may provide the opportunity to the decision-making system to re-optimize its objectives. The approach provided here is the latter and ensures that prediction of flexibility at PV level is provided in the short term.

8.2.6.2 Background

Regarding PV generation forecasting models per se, the relevant background literature review has been reported in the respective section (3.20) of D4.2. In this section the technology background focuses particularly on generation (from PV) flexibility models.

The idea of power system flexibility has been introduced to describe the modification of generation injection and/or consumption patterns in reaction to an external signal (price signal or activation) in order to provide a service within the energy system. This generally encompasses the extent and speed with which generation or consumption levels can be changed. Historically, DER only marginally contributed to power system flexibility, but the provision of flexibility from DER can also contribute to both global/market wide and local flexibility challenges. There are



however several challenges for efficient market participation of flexible DER, and new designs or business models are likely to be needed to facilitate flexible DER².

8.2.6.3 SYNERGY Implementation Details

8.2.6.3.1 *Input Data*

The dataset utilised to train and evaluate the generation prediction model that is embedded in the generation flexibility prediction model comprises hourly day-ahead Numerical Weather Prediction (NWP) data for a specific location in Greece where a 500KW PV plant is located, as well as actual hourly production data (kWh) from the same PV plant for the same period of time.

As in the case of the “Demand Flexibility Prediction model at Building Level”, the weather forecasted data were retrieved from OpenWeatherMap³. A number of trials were carried out in the training phase of the model with respect to the number and type of weather metrics to be considered as inputs to the model. It was found that the combination of weather metrics providing the best forecasting accuracy for the model includes Ultra-Violet Radiation Index - UVI, temperature (°C) and cloud coverage (%).

The granularity of training, evaluation and test data is hourly while the training set covers a period of several months resulting in a training dataset comprising of 1500 entries of the four aforementioned features (uvi, temperature, clouds, energy yield).

8.2.6.4 Approach

Step 1: Feature pre-processing

The model includes the following pre-processing steps that apply on the dataset:

- Data normalization (scaling). A MinMax scaler was utilized.
- Dataset reshaping in an appropriate 3D matrix which is a hard requirement in LSTM models.

Step 2: Model training

²

<https://ec.europa.eu/energy/sites/default/files/documents/5469759000%20Effective%20integration%20of%20DER%20Final%20ver%2026%20April%202015.pdf>

³ <https://openweathermap.org/>



The embedded PV generation prediction model is a recurrent neural network with long-short term memory units (RNN - LSTM). Keras library was utilized to define the sequential class of the model comprising of an LSTM layer with a number of hidden cells followed by a Dense output layer. The model uses batches of 24 input-output pairs while the same activation function is applied on both LSTM and Dense layers.

Step 3: Model performance evaluation

The performance of the generation flexibility prediction model coincides with the performance of the embedded PV generation prediction model. The initially available dataset was split such that 80% was used for training and the rest 20% for the evaluation of the model. The evaluation pipeline fed the model with 24-hour ahead weather forecasts, namely UVI, temperature and clouds coverage -unseen by the model during the training stage - which in turn returned the predicted PV generation. The model is compiled using an “Adam” optimizer which outperformed previous trials with stochastic gradient descent (SGD) optimizer. The metrics utilized to evaluate the accuracy of the model were the mean absolute error (MAE), and “Accuracy”, The reason that MAE was preferred over MSE metric is that it offers the advantage of preventing outliers when PV production is equal or close to zero (division by zero problem).

8.2.6.5 Technology

The day-ahead PV generation prediction model that lays at the core of the generation flexibility prediction solution, is a Python model, developed and trained outside the SYNERGY environment. The basic libraries utilized for the training of the model are: Tensorflow-Keras, Scikit-learn, Pandas and Numpy.

The versions of the libraries used for the process training are aligned to the model registration guidelines. The trained model was saved in “h5” format and registered in the SYNERGY Platform.

After registering the aforementioned model in the SYNERGY Platform, a pipeline was created through which the aforementioned model is applied in the context of a more complex workflow -as presented in the following figure- that enables the prediction of short term (day-ahead) generation flexibility at DER level, through the comparison of an early generation prediction with an updated one (closer to realization).

The application pipeline comprises two, initially independent flows, each one utilizing different input datasets but both primarily resulting in the day-ahead generation prediction of the PV installation. The critical difference between the two flows is that the one is based on early day-



ahead weather forecasted input parameters, retrieved from an external weather service provider at 09:00 am of previous day (D-1), while the other is based on an updated day-ahead weather forecast, retrieved at 21:00 pm of D-1, from the same source. Prior to the input datasets being ready to feed the generation prediction models they enter a pre-processing stage including normalization and other vetting processes in order to comply with the generation prediction models’ requirements. Ultimately, the generation flexibility model’s output is the signed flexibility in kW (upwards, downwards), calculated as a result from the comparison of the early generation prediction versus the updated one (closer to realization).

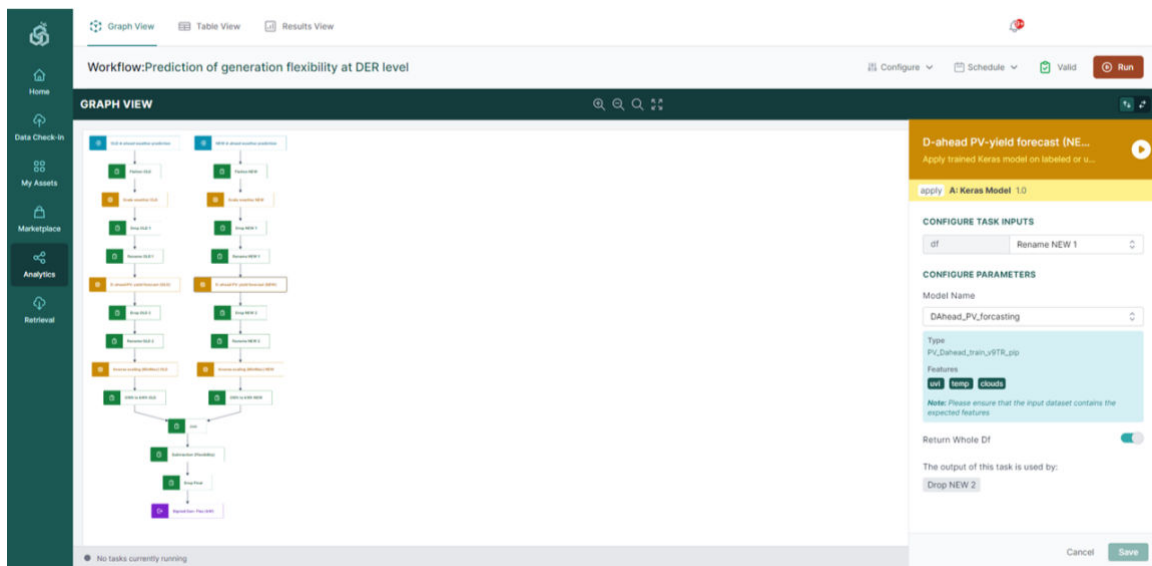


Figure 119: Model Application Workflow

8.2.6.6 License & Availability in the SYNERGY Platform

The trained model is available in the SYNERGY Platform, under the name “D-ahead PV forecasting”. It can be accessed through the SYNERGY marketplace and used in an analytics workflow in a similar way to the one presented in Error! Reference source not found., under the terms of the established contract between the data asset provider (VERD) and consumer.

The aforementioned workflow is also accessible in the SYNERGY platform by the name “Prediction of generation flexibility at DER level” to facilitate users who obtain the aforementioned model. The trained model and the pipeline are provided with a confidential license by their provider, VERD, that is the owner of the model’s intellectual property rights.

8.2.6.7 Assumptions and Limitations

Although the generation prediction model that lays at the core of the generation flexibility model under discussion, is found to provide satisfactory results when used for generation forecasting of PV plants located in various areas around Greece, its efficiency is expected to be limited when used in a wider geographical range due to the localized nature of the training process. Diverse PV panels efficiency apparent in other plants will also not be reflected in the output of the currently trained model. In order to address this issue, a new training dataset must be available through SYNERGY in order to retrain the model taking into account the special characteristics of the user’s PV plant and make it available through SYNERGY. The description of the training dataset will be available through the platform, which can help interested stakeholders understand whether their intended usage will be a good fit for the already trained model. At a later stage of the project, the possibility of making the training pipeline available in the SYNERGY Platform will be investigated.

8.2.7 **Anomaly detection in energy demand at building level**

Scope	Industrial Analytics
Problem	Demand Forecasting
Question	IND.I.12 Are there any anomalies/outliers detected in the energy demand at building level?

8.2.7.1 Description

Large amounts of data are being produced everyday by smart sensors installed in buildings. If leveraged properly, that data could assist end-users, energy producers and utility companies in detecting anomalous power consumption and understanding the causes of each anomaly. The applications of anomaly detection of energy consumption are no longer limited to energy efficiency, but they are finding themselves in various novel application contexts (Himeur et al., 2021). In particular, they could be used for detecting (i) abnormal consumption behaviors, (ii) faulty appliances, (iii) occupancy information, (iv) non-technical losses, and (v) at-home elderly monitoring. In addition, the same anomaly detection system, can be used for multiple applications without the need for installing other systems (e.g. to detect occupancy or non-technical losses). Therefore, this could effectively reduce the hardware implementation costs and decrease the complexity of installed systems.



In this section, an algorithm for anomaly detection in the energy demand at grid level is proposed. In particular, the proposed algorithm utilizes energy consumption data, which are collected hourly from a SMART grid, while it also utilizes weather (e.g. temperature, humidity, etc.) and calendar data (e.g. public holidays).

8.2.7.2 Background

Anomaly detection refers to the process of detecting abnormal events that do not conform to expected patterns. Broadly, depending on their types, anomalies can be classified as point, contextual or collective anomalies (Chandola et al., 2009). If a data instance is anomalous compared to the rest of the data, then it is referred to as a point anomaly (e.g., a daily lighting energy consumption value might be anomalous compared to previous values). If a data instance is normal in one context but anomalous in another, then it is referred to as a contextual anomaly (e.g., an hourly school lighting consumption value might be anomalous on weekends when there are no classes but not on weekdays). If a group of related data instances is anomalous in comparison to the rest of the dataset, then it is referred to as a collective anomaly. Individually, these values might not be anomalous, but collectively they represent an anomalous occurrence. For instance, the individual values of a daily profile of heating, ventilating, and air conditioning consumption data might be normal compared to previous recorded values, but collectively, the daily profile might represent an anomalous consumption pattern.

Anomaly detection of energy demand has been examined by several researchers using both unsupervised and supervised detection algorithms. In particular, Jakkula & Cook (2011) utilized one-class support vector machine (OCSVM) to identify the smallest hypersphere encompassing all the power observations, while Ghori et al. (2020) proposed one-class random forest (OCRF) to identify abnormal consumption when labelled data are absent. Henriques et al. (2020) and Yeckle et al. (2018) utilized k-means for separating anomalous and normal events in highly coherent clusters. Wun et al. (2019) used isolation forest with split-selection criterion algorithm to check if the end-user's electricity consumption is anomalous or normal. In addition, various dimensionality reduction techniques were explored to classify power data as normal or abnormal, such as principal component analysis, linear discriminant analysis (Valko et al., 2011) and quadratic discriminant analysis (Naveen et al., 2016).

Yuan et al. (2015) and Himeur et al. (2021) used autoencoder to detect abnormal energy usage, while Weng et al. (2018) and Wang et al. (2019) merged autoencoder and long short-term memory (LSTM) neural networks to identify abnormalities in unbalanced and power



consumption datasets. da Silva et al. (2019) used Recurrent neural network (RNN) to predict the anomalies occurring during energy usage and distinguish them from deviations emerging from seasonality, weather and holiday dependencies. Also, convolutional neural networks (CNN) have demonstrated its effectiveness for detecting abnormalities in time-series data (Li et al., 2019; Zheng et al., 2017).

Jakkula et al. (2010) and Liu et al. (2016) deployed statistical algorithms to identify the anomalies via the identification of extremes based on the standard deviation. Sial et al. (2021) proposed KNN based heuristics to detect abnormal power consumption, while Mulongo et al. (2020) investigated the performance of KNN against other machine learning classifiers to identify abnormal power observations. Depuru et al. (2011) and Korba et al. (2019) deployed SVM to detect abnormalities due to energy theft attacks. Kammerer et al. (2019) used a decision tree solution to learn energy consumption anomalies. Touzani et al. (2018) and Tama et al. (2019) utilized a GBM algorithm to model power usage. Finally, Primartha & Tama (2017) used a random forest classifier to detect anomalies while respecting the performance measure related to the accuracy and false alarm rates.

8.2.7.3 SYNERGY Implementation Details

8.2.7.3.1 *Input Data*

The dataset used to train and evaluate the model was retrieved from the private database of the University of Cyprus. The dataset (7,560 rows) contains the hourly energy consumption between the period 01/04/2019 – 29/02/2020 of 1 smart grid from University of Cyprus. For expanding the features of the dataset, historical weather data were retrieved from World Weather Online⁴ using a free trial version of the tool. The World Weather Online contains historical weather data from 1st July 2008 onwards for several cities and its data are accessible through an API.

Below we provide the expected data columns of the dataset along with their description and format:

- G_UCY: The hourly energy consumption of the smart grid (float)

⁴ <https://www.worldweatheronline.com/developer/>



- weekday: The day of the week (int with range of values [0-6])
- day_of_month: The day of the month (int with range of values [0/31])
- hour: int, range of values [0-23]
- isHoliday: Whether the specific day is a public holiday or not (binary with range of values [0/1])
- tempC: Temperature in degrees Celsius (int)
- humidity: Humidity in percentage (int, range of values [0-100])
- windspeedKmph: Wind speed in kilometers per hour (int)
- precipMM: Precipitation in millimeters (float)
- cloudcover: Cloud cover amount in percentage (int with range of values [0-100])

8.2.7.3.2 Approach

Step 1: Feature pre-processing

- **Detect multicollinearity:** Both correlation coefficient and Variance Inflation Factor (VIF) are used to verify multicollinearity. Dohoo et al. (1997) argued that multicollinearity is certain at the 0.9 level of a correlation coefficient or higher. On the other hand, the VIF values for included variables should be below 10. If any (correlation coefficient or VIF) critical value is higher than described above, then an independent variable is removed from our set.
- **Clean Data:** Remove records with incorrect energy consumption values due to system failures and/or maintenance.

Step 2: Anomaly detection and labelling

- **Point Anomalies:** For detecting point anomalies (i.e., extreme values) in the energy consumption, we used an exponentially weighted moving average of the energy consumption feature with a seven-day moving window. A short moving window adapts quickly to changing values and seasonality effects. In addition, we defined as anomalies all values that differ by more than two standard deviations from the mean.
- **Contextual anomalies:** For detecting contextual anomalies (e.g., an hourly consumption value that is anomalous on weekends when there are no classes), we utilized the



Isolation Forest model using all the features available on the dataset. Specifically, the Isolation Forest model was trained using the following parameters:

- contamination=.01
- n_estimators=500
- The records that were either point or contextual anomalies were labelled as 1 (anomaly), while the rest of the records were labelled as 0 (normal).

Step 3: Model training

- For detecting the anomalies in the energy demand, we apply a one-class support vector machine (OCSVM) model. For training the model, all the anomalies were removed and only the inliers were fitted into the model. The developed model trained with the following parameters:
 - kernel='linear'
 - nu=0.05
- Using one-class support vector machine, we apply the prediction function to each of the records in the test set, which returns -1 for outliers/anomalies and 1 for inliers/normal.

Step 4: Model performance evaluation

- The dataset was split into training and testing set, corresponding to 75% and 25% of the dataset respectively. Specifically, the training set consisted of 5,670 samples and the test set of 1,890 samples. Finally, precision, recall and F1 score were used in order to select the right model parameters and evaluate the model.

8.2.7.3.3 Technology

The model was developed and trained outside the SYNERGY environment using the Scikit-learn library. Then, the trained model was exported as a pickle file and registered in the SYNERGY Platform. The versions of the libraries used for training are the ones provided in the model registration guidelines. After registering the trained model in the SYNERGY platform, a pipeline was created for applying the trained model.



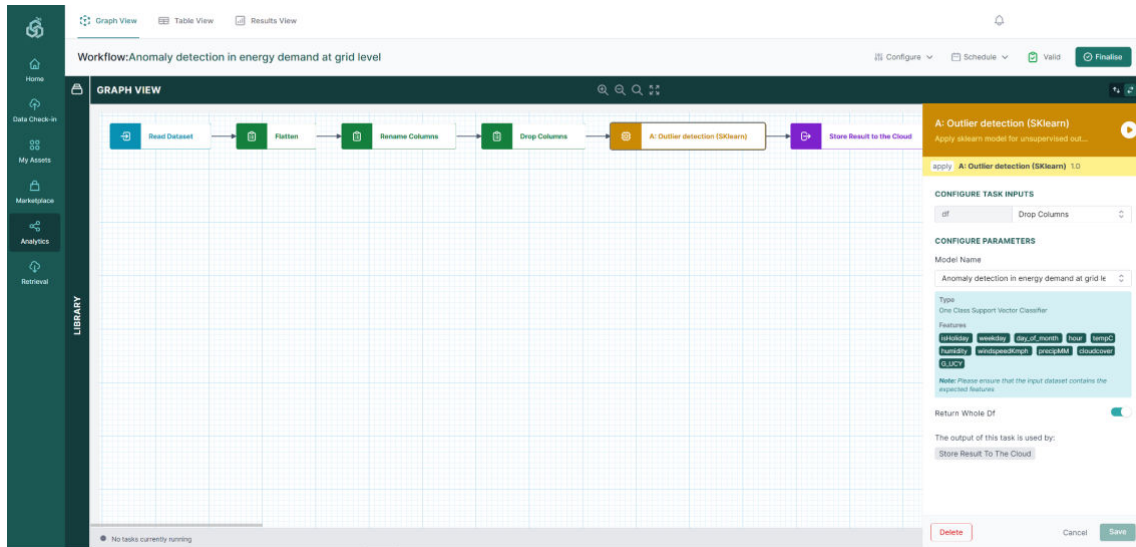


Figure 120: Model application workflow

8.2.7.3.4 License & Availability in the SYNERGY Platform

The trained model is available in the SYNERGY Platform with the name “Anomaly detection in energy demand at grid level”. It can be acquired through the SYNERGY marketplace and used in any pipeline under the terms of the established contract between the data asset provider (UCY) and the data asset consumer.

The model is provided with a confidential license by its provider, UCY, that is the owner of the model’s intellectual property rights.

8.2.7.3.5 Assumptions and Limitations

First, we assume that the anomalous data were identified accurately by the labelling process. As future enhancement, the anomalies will be validated and corrected by human experts from University of Cyprus. Second, due to the nature of unsupervised learning, the computational complexity of the training time may be huge when the training set is large. Third, the pre-trained model in order to be used must be presented with the exact input format used in the training process. (e.g., energy consumption data and weather data need to be collected and aggregated per hour).

8.2.8 Outlier detection in energy demand

Scope	Industrial Analytics
--------------	----------------------



Problem	Demand Forecasting
Question	IND.I.15 - Are there any anomalies/outliers detected in the energy demand at portfolio level?

8.2.8.1 Description

Modern power systems are evolving in a more sustainable path. The load demand for domestic electrical energy is gradually increasing as the number of household appliances and electric cars increases. Statistics show that residences and commercial buildings account for three-fifths of global electricity use. The power system has grown in complexity and intelligence, and more modern information transmission technology has been implemented, making grid processing more convenient and secure. Moreover, electric energy consumption in everyday living is also difficult and variable. Electric energy usage, for example, may vary significantly depending on the season, and consumption on working days and working days will fluctuate. At the same time, there will be anomalies in the electrical load, such as forgetting to turn off electrical appliances, failure of electrical appliances and even the theft of electricity, and so on, resulting in a much larger electrical demand than typical. As a result, detecting unusual consumption data is critical. Abnormal detection can enhance abnormal electric energy consumption to achieve energy savings, remind users to discover malfunctioning electrical appliances or modify bad electricity usage patterns, lower users’ energy consumption expenses, and promote electricity consumption safety awareness. The most crucial factor is that you can locate the source of the power theft. According to the survey, power theft accounts for about half of the energy lost in some developing countries, and anomaly detection technologies can successfully combat this scenario.

Anomaly detection, as the name suggests, is the method of recognizing data that differs from the usual. Anomalies in data are situations that do not follow the specified usual behaviour pattern.

8.2.8.2 Background

(Papastefanopoulos et al., 2021) and (TowardsDataScience, 2020) explain some different approaches to afford the outlier detection problem. These different approaches are divided as follow:



- **Unsupervised detection:** It aims at detecting formerly unknown rare consumption observations or patterns without using any a priori knowledge of these observations. Generally, this kind of detection assumes that the amount of anomaly patterns to the overall consumption data is small, i.e. less than 20%. Since the abnormalities represent the outliers that are unknown to the consumer at the training stage, detecting anomalous consumption is reduced to the modelling of normal consumption behaviour in the large majority of cases, in addition to the definition of specific measurements in this space with the aim of classifying consumption observations as abnormal or normal. Unsupervised techniques are mainly built on *clustering, one-class learning, dimensionality reduction algorithms* and NN Encoders-Decoders.
- **Supervised detection:** Supervised anomaly detection in energy consumption necessitates training the machine learning classifiers (binary or multi-class) using annotated datasets, where both normal and abnormal power consumption is labelled. Although supervised anomaly detection can achieve high-accuracy identification results as demonstrated in academic frameworks, its adoption in the real world is still limited compared to unsupervised methods, due to the absence of power consumption annotated datasets. Supervised techniques are mainly built on *Neural Networks (LSTMs, CNNs...), Support Vector Machine, K-nearest neighbours*.

8.2.8.3 SYNERGY Implementation Details

8.2.8.3.1 *Input Data*

The model input consists of a dataset with two columns. The first one was the datetime column and the second one corresponds to the energy demand value. Thanks, of the simplicity of the data, this approach could be applied to any energy demand data, but in this approach the data is at area level.

In this case, the data has been obtained from Kaggle web where there are a lot of public datasets. The data registers could be separated by 15 minutes, hours, days or months. Ideally the data should be separated by hours and contain at least 1 year of historical data.



8.2.8.3.2 Approach

Before starting with the algorithm implementation, it is necessary to remark a necessary data pre-processing to give more information to the data.

The following attributes were added to the initial energy demand value column:

- **Hour** (Related hour)
- **Month** (Related month)
- **Weekday** (Related day number of the week)

Once the data were pre-processed and normalized (*energy demand value*), an Encoder-Decoder topology was used to face the problem of unsupervised detection. This topology follows the next structure:

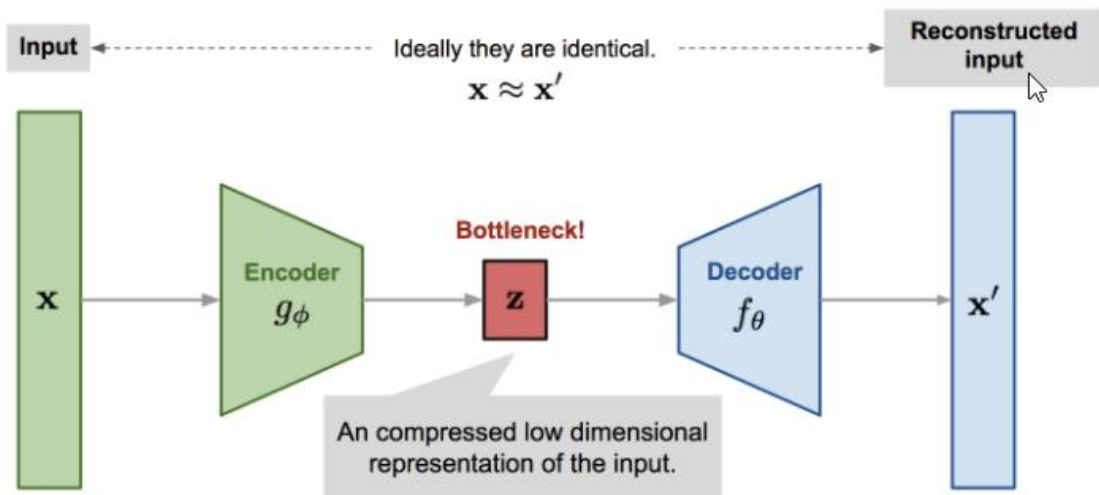


Figure 121: Encoder-Decoder Topology (Source:<https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html>)

Where the encoder was a down-sampling MLP (Multi-layer perceptron) and the decoder was an up-sampling MLP.

When the neural network was trained, it was able to give an error for each sample reconstruction and this error was the point used to detect if a sample was an anomaly or not.

However, this error was not enough to detect an anomaly, because a criterion was needed to use this error. To obtain this criterion, a threshold was to be defined using a normal distribution of the training errors, as it is shown below:

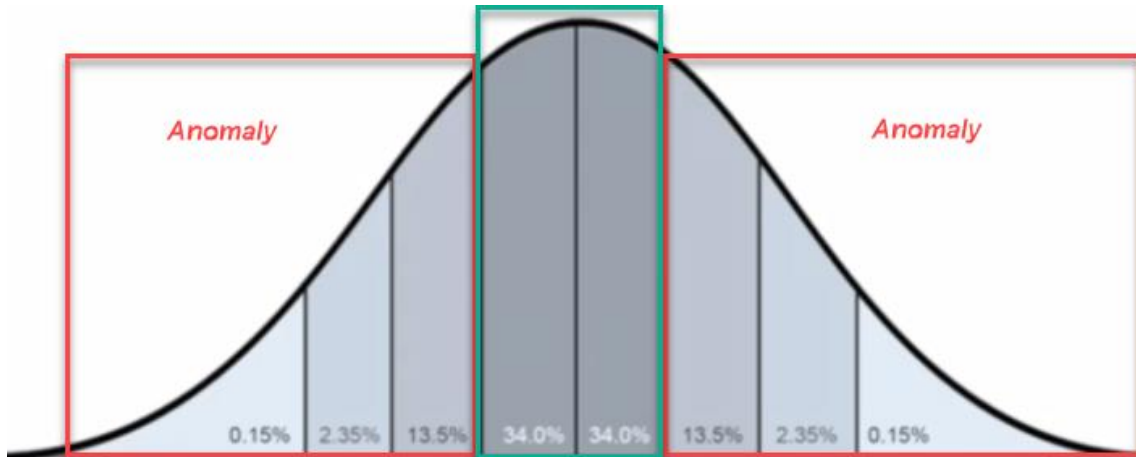


Figure 122: Anomaly threshold setting

Where all samples with an error reconstruction bigger than errors in the centre of the curve were labelled as anomaly.

8.2.8.3.3 Technology

This problem requires to create two workflows, where the first one is necessary to train the “Encoder-Decoder” model and the second one uses this trained model to receive the reconstruction error, to be used with the pre-calculated threshold to label anomalies.

Training stage

This stage contains two steps, the first one was used to train a Min-Max scaler with the energy demand value, as depicted in Figure 123

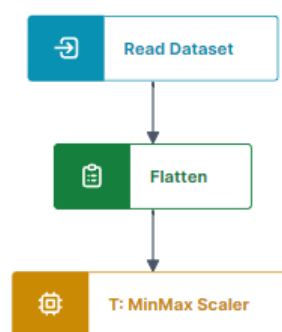


Figure 123: MinMax Scaler Training

And the second one is used to read data and manage it adding the required information to create the model features. Once the columns were added, the pretrained scaler was used to normalize

the energy demand value and after that an encoder-decoder was trained, as shown in *Figure 124*:

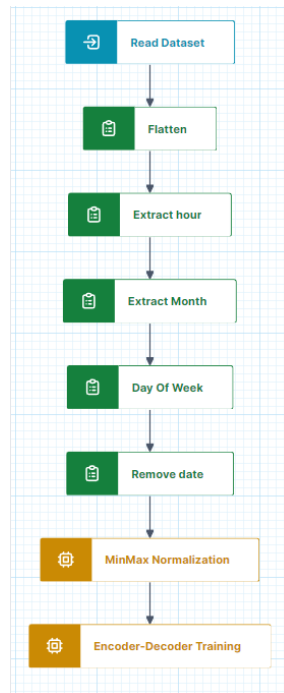


Figure 124: MinMax scaler application and encoder-decoder training

Used libraries:

- Tensorflow
- Scikit-learn
- Pandas

8.2.8.3.4 License & Availability in the SYNERGY Platform

ETRA I+D is the owner of all intellectual property rights of this model. All rights are reserved.

The model is available in the SYNERGY Platform as a block to be included in custom analytics workflows upon agreement with ETRA I+D.

8.2.8.3.5 Assumptions and Limitations

This approach assumes that the model will be trained in each execution to adapt it to new input data. This way, any dependence between the model and the input data is avoided. The input data must contain a column with the date and the hour and another column with the consumption.



8.2.9 Prediction of peak and average energy generation at portfolio level in specific TSO/DSO areas (over a year)

Scope	Industrial Analytics
Problem	Generation Forecasting
Question	IND.II.8 - What is the expected peak and average energy generation at portfolio level in specific TSO/DSO areas in the next hour?

8.2.9.1 Description

For systems in which, distributed generation penetration is high, the accurate prediction of expected peak and average of energy production is a very important asset. Planning can be designed with more efficient analysis when portfolio level statistics are forecasted. Also, system operators have more system indices available. Prediction of penetration is valuable and can contribute in success of meeting demand in the most efficient way. A common problem in order to achieve a secure and uninterrupted operation of a power system with minimum energy waste, is the temporal mismatch between the peak of energy generated from RES and the peak of energy demand in a day. This problem intensifies as the penetration of RES is increased. Often, this temporal mismatch between the generation and demand peaks is covered by incorporating expensive forms of energy sources with a high impact on the environment. Therefore, an accurate energy generation peak and average prediction model can provide essential information to power system operators, in order to make optimal decisions while it aids in performing system planning.

8.2.9.2 Background

Energy generation forecasting is time series forecasting problem. It can be classified as multivariate because there is a set of parameters that affect the generation. The forecasting procedure is mainly addressed as a data driven regression problem. In this particular case of forecast, the most widely used models, with highest accuracy, are machine learning and Deep learning ones, such as Artificial Neural networks, Support vector machines and Random Forests (Wang et al., 2019). Also, autoregressive integrated moving average models (arima,sarimax) are being used in an ensemble method (Natarajan et al., 2019). The idea behind these methods, is to use a data-driven model, in training process where the model will define a function of weather



variables and previous observed values of a generation time-series in order to predict the energy generated at portfolio level (Gireeshma et al., 2019). The model aims to approximate the generation systems' behaviour, transforming it into base units. The model is performing a long-term energy forecast in order to statistically address the mean and peak values per month. After the training process, the data-driven model will be using the appropriate input data and estimate the energy generation of the whole portfolio under study. As the volume of available data increases exponentially, more advanced methods will be designed based on Deep Learning.

8.2.9.3 SYNERGY Implementation Details

8.2.9.3.1 *Input Data*

The energy generation peak/mean prediction models require as input one dataset. The dataset contains previously measured historical values of the generation, for a specific time period. Also, in combination with the energy measurements, nominal values of the installed capacity under study must be passed. In addition, historical weather measurements can be used to widen the analysis scenario.

8.2.9.3.2 *Approach*

The energy generation forecasting is addressed as a time series forecasting problem. The time span of one year, consists of approaching the processes as a long-term forecast.

Once the data pre-processing and all the variables are defined, a deep learning architecture is designed and trained. For the specific task a fully connected neural network activation output layer is used, in combination with a sarimax model. The best performing model is then chosen and used to generate the forecast.

8.2.9.3.3 *Technology*

All designed models use the following libraries (in the versions supported by the SYNERGY platform):

- Tensorflow
- Keras
- Scikit-Learn
- Pandas



- NumPy
- Statsmodels

After registering the trained model in the platform, a workflow is created, in which the input data are processed in order to be in the correct form. Everything needed is uploaded within the SYNERGY platform. A publicly available dataset sourced from ENTSO-E⁵ has been used. After data cleansing and reconstruction, it has been uploaded and is available to be used in the Assets section.

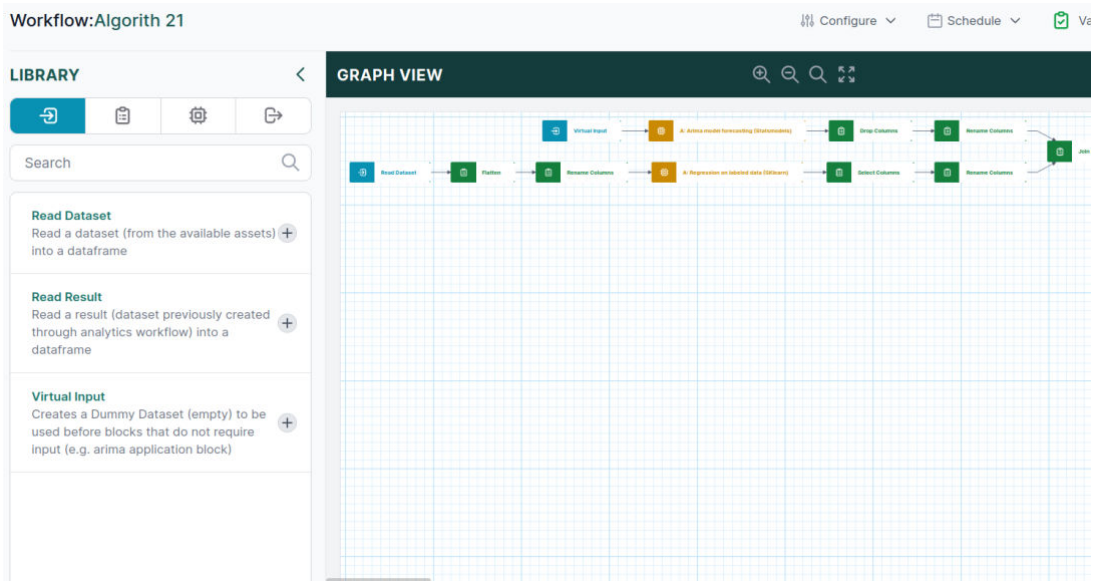


Figure 125: Application workflow (partial view)

8.2.9.3.4 License & Availability in the SYNERGY Platform

All models are registered and available in the SYNERGY platform as a service by the respective data asset provider (ICCS) and can be shared and applied through the platform’s functionalities.

8.2.9.3.5 Assumptions and Limitations

The energy generation prediction tasks are considered as time series regression forecasting tasks. The quality of the historical data measurements of previously generated energy, in combination with a large past time horizon enable the model to perform. Lack of data can lead

⁵ ENTSO-E: European association for the cooperation of transmission system operators (TSOs) for electricity website: <https://www.entsoe.eu/>

to poor long-term predictions. After the training procedure, the models in order to be used must be presented with the exact input format used in the training process.

8.2.10 Prediction of storage flexibility

Scope	Industrial Analytics
Problem	Flexibility Forecasting
Question	IND.IV.2 - What is the expected (very) short-term storage flexibility at building level under a self-consumption framework?

8.2.10.1 Description

Self-consumption can be defined as the share of the total PV production directly consumed by the PV system owner. It can be further increased when combined with demand side management or when energy storage technology is leveraged, the latter being able to provide an increase of up to 24% for residential consumers whose self-consumption rates range between 20%-40% (Luthander et al., 2015). Increased self-consumption offers economic benefits to prosumers who can also have better control over their electricity bill, and contributes towards lowering the stress on the electricity distribution grid.

Energy storage has thus emerged as an attractive solution enabling storage during periods of high generation to be later used during periods of high demand (Vieira et al., 2017). Combined with the decreasing battery costs, consumers’ interest in increasing PV self-consumption by using battery systems is also increasing (Pena-Bello et al., 2019).

The setting that is of interest in the current analysis thus refers to residential PV systems that are equipped with batteries to optimize the utilization of the energy generated by the PV in a way that maximises the onsite load covered either directly by the PV or through the battery (storage). The aim here is to provide an initial approach to compute the energy flexibility that is in this way offered by the storage system.

8.2.10.2 Background

It should be noted that the battery flexibility in this context depends on the demand and generation, but these are not the only factors that affect how the decision over battery charge



and discharge is made. Indicatively in (Pena-Bello et al., 2019), which explores how different battery technologies operate in these systems, the optimal battery energy storage system control strategy is shown to be application-specific and in particular the following four applications are examined:

- (a) PV self-consumption with the main driver being the price difference between the electricity imported from and exported to the grid
- (b) Avoidance of PV curtailment which is of interest in areas with high PV penetration, where grid stability issues may emerge
- (c) Demand load-shifting again for tariff related reasons, e.g when prices are lower in off-peak periods, so battery discharges are scheduled for when prices are high, i.e. in peak periods.
- (d) Demand-peak shaving used to mitigate electricity peaks

The authors in (Vieira et al., 2017) present the design of an energy storage system for residential buildings and propose two control strategies for the battery charge and discharge process, one for cases with higher generation and one for cases when demand exceeds generation. Their aim is to convert the building into a zero-energy building and they apply their system on simulated data to test various solar radiation and demand conditions, achieving a reduction on the annual energy bill of 87.2%, as well as reduction of power flows with the grid.

(Martins et al., 2016) presents an advanced controller that uses PV generation and the demand profiles to define an improved dispatch operation for photovoltaic/energy storage system in a group of households using linear programming to solve the formulated optimization problem. In (Luthander et al., 2015), the size of the battery storage normalized by the size of the PV system emerges as an important factor in increasing self-consumption. At the same time, the authors identify a need for more research in the field, in particular in terms of comparative studies that will offer a better understanding of the potential of such systems.

8.2.10.3 SYNERGY Implementation Details

8.2.10.3.1 *Input Data*

The input data used in the particular analytics solution are the following:

- Short-term (hours ahead) PV generation forecast for the PV of the residential building of interest
- Short-term (hours ahead) demand forecasting for the same residential building
- The battery capacity and the minimum state of charge



8.2.10.3.2 Approach

In the current solution, the only driver of the battery charge and discharge decisions is the difference between expected generation and demand, i.e. the simplest self-consumption setting. Assuming “perfect” forecasts, the battery flexibility is calculated as the difference between these two, taking into consideration the current charging stage of the battery. It should be noted that the two forecasts are considered inputs and are thus not part of the implemented solution, which is not affected by the way the forecasts are generated, as long as they are available.

8.2.10.3.3 Technology

The approach is implemented as a pipeline within the SYNERGY platform, as shown in *Figure 126*, and hence the versions of the libraries that are used are the ones provided by the platform.

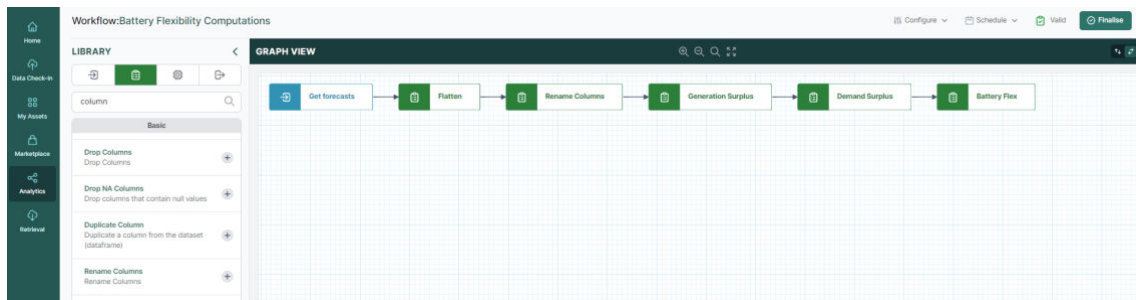


Figure 126: Flexibility computation workflow

8.2.10.3.4 License & Availability in the SYNERGY Platform

As explained, the current implementation does not implement and train any new machine learning models. Instead, it leverages other solutions of the analytics catalogue in order to obtain the generation and demand forecasts and adds an additional business layer in the way these are combined with the battery information to extract the storage flexibility. The created workflow is provided with a confidential license by its provider, Suite5, that is the IPR owner of the specific workflow.

8.2.10.3.5 Assumptions and Limitations

The generation and demand forecasts are of paramount importance for the computation of the expected battery flexibility, therefore the implemented solution is as accurate as the underlying models that will be selected as inputs. It should also be stressed that the current solution does not consider all attributes of these systems, such as battery cycle life, calendar life, voltage

limitations and charging/discharging efficiency aspects. Finally, as explained, the scenario explored here is the simplest case of self-consumption, but solutions that implement more advanced strategies will be investigated in the future once the relevant demo data are available.

8.2.11 Malfunction Duration Prediction in PV energy generation

Scope	Industrial Analytics
Problem	Predictive and Preventive Maintenance
Question	IND.V.2 - How long will a malfunction/inefficiency at generation asset level last?

8.2.11.1 Description

When a malfunction occurs, it is important for energy providers to know exactly how long it takes for the PV plant to return to normal energy generation level. But since usually there is no identified or registered status class of the activity that started the PV malfunctioning, the first action to be taken needs to be the detection of what triggered the process. Consequently, to predict the duration of malfunctions in the performance of PV assets, also a process of malfunction detection is required. For that, the model needs to be trained in two steps:(a) anomaly detection: to detect anomalies and malfunction in the energy generation level of PV assets, (b) duration: to predict since how long the anomaly exists.

From a high level and generic perspective, time series anomaly detection can be done in two ways: (a) building a predictive model using the historical data to have insights of the overall common trend, seasonal or cyclic pattern of the time series data such as ARIMA, Regression or LSTM; (b) building a predictive model using Unsupervised Cluster-based Approaches such as KNN, Kmeans, DBSCAN, Isolation Forest. This baseline analytics uses KNN, that although being a supervised ML algorithm, when it comes to anomaly detection it takes an unsupervised approach.

8.2.11.2 Background

The need to detect anomalies/inefficiency in PV energy assets has been explored in many academic studies. In usual consumer practices the PV systems are set with a motto of “Set and forget”. But during a large period of time, if not kept in check, the performance of PV systems degrades, and anomalies get undetected which leads to reduced energy generation.



(Bosman et al., 2020) provided a brief review of present approaches for predictive maintenance of PV assets. The four approaches mentioned by the authors are as follows:

- **Manual Diagnostics:** the least expensive and the lowest detection accuracy approach including qualitative (visual inspection of the system and individual components and infrared thermography of PV panels) and quantitative approaches (I-V curve and insulation resistance analysis of PV panels).
- **Failure Mode and Effects Analysis (FMEA):** moderately expensive approach, offers a medium amount of detection accuracy. It is a semi-qualitative method used to prevent failures and analyse the risks of a process, by identifying causes and effects on the system to determine the actions used to prevent failure.
- **Machine Learning & Forecasting:** moderately expensive approach, that offers a medium amount of detection accuracy. It uses Machine Learning algorithms to estimate energy generation and find critical factors affecting it, unveiling temporal patterns in the weather data, etc.
- **Real-Time Sensors:** most expensive approach but offers the highest detection accuracy that uses sensors for the purpose of tracking, where the PV panel automatically orients according to the sun's direction.

Even though there is a lot of work on detecting and predicting anomaly events in PV systems, the studies on predicting the duration of these anomalies are limited. (Hu, 2012) proposed several classifications of anomaly types and proposed an extensible framework to detect these anomalies at a time series-based estimation of how long these anomalies last. A time series that influences an anomaly is described using the following intervention model: $Y_t = A_t + z_t$. In this intervention model z_t describes an ARIMA time series without any intervention while A_t is used to describe the intervention due to anomalies. A_t is the core part of this model because it contains an occurrence function which describes the start and end time of anomalies. Y_t is then the observed time series.

8.2.11.3 SYNERGY Implementation Details

8.2.11.3.1 *Input Data*

To train and evaluate the PV Malfunction Duration Prediction model, an online open-source dataset “Solar Power Generation Data” was downloaded from Kaggle



(<https://www.kaggle.com/anikannal/solar-power-generation-data>). This is a data source that has been gathered at a solar power plant in India over a period of 34 days. The observations were recorded at 15-minute intervals. It consists of the following files:

- The power generation datasets are gathered at the inverter level - each inverter has multiple lines of solar panels attached to it.
- The sensor data are gathered at a plant level - single array of weather sensors optimally placed at the plant.

After aggregating the power generation dataset, to have the generated energy at the plant level, the data tables were joined, data were cleaned thus removing valueless features, a dataset of 3158 entries, with the following four variables was obtained:

- DC power: Amount of DC power generated by the PV plant during the 15 minutes interval (Units – kW).
- Ambient temperature: the ambient temperature at the plant recorded at 15 minutes intervals.
- Module temperature: the temperature of the module attached to the sensor panel recorded at 15 minutes intervals.
- Irradiation: Amount of irradiation for the 15 minutes interval.

8.2.11.3.2 Approach

In the baseline model, the DC generation of the PV plant is the considered target. After preparing the data, the final model was created in two steps, as detailed before: (a) anomaly detection model: to detect anomalies and malfunctions in the energy generation, (b) duration prediction model: to predict since how long the malfunction exists.

Step 1: Feature pre-processing

In order to clean and prepare the data that will be used to train the model, the following pre-processing steps are performed:

- Data are aggregated to have enough information about the energy generated at the plant level
- Data tables available are joined for namely Power Generation dataset and Sensor gathered dataset



- Data are cleaned (the features with missing values that appear after joining are quantified)
- Feature selection is performed (valueless features such as Planet ID, source key, etc are removed)
- Features are normalised

Step2: Malfunction detection and labelling

- After preparing the data, a malfunction detection model is created on the DC power to detect anomalies on the energy generation of the PV plant. The model is a KNN anomaly detection, that has the DC power variable (energy generation) as an input. The output is a binary value for each timestamp, considering if it is an anomaly or not (the output is 0 if it is not an anomaly and 1 if it is).
- Since the final model is intended to predict the duration of the malfunctions, each timestamp must be labelled based on its duration. Four status/categories are considered: (1) 0 minutes (normal generation); (2) [0-15] minutes; (3) [15-30] minutes and (4) >30 minutes

Step 3: Random Forest Model and label prediction

In this step, the anomaly is framed according to a classification problem. Here a Random Forest model is used, which is trained with the following parameters:

- Input variables: including ambient temperature, module temperature and irradiation
- Label: malfunction duration label

A 10-fold cross validation was performed to evaluate the performance of the model. The result of the validation showed that the pretrained model can predict the malfunction duration with a high accuracy.

8.2.11.3.3 Technology

The PV malfunction duration prediction model was developed and trained outside the SYNERGY Platform in two model training steps including malfunction detection and malfunction duration prediction using the Scikit-learn (SKlearn) library. Afterwards it was packed into the pickle format and registered in the SYNERGY Platform. It uses the available libraries and versions provided by the platform being the ones provided in the model registration guidelines.



After that, a pipeline was created to use the pre-trained model and show how it can be applied. The created pipeline is shown in Figure 127:

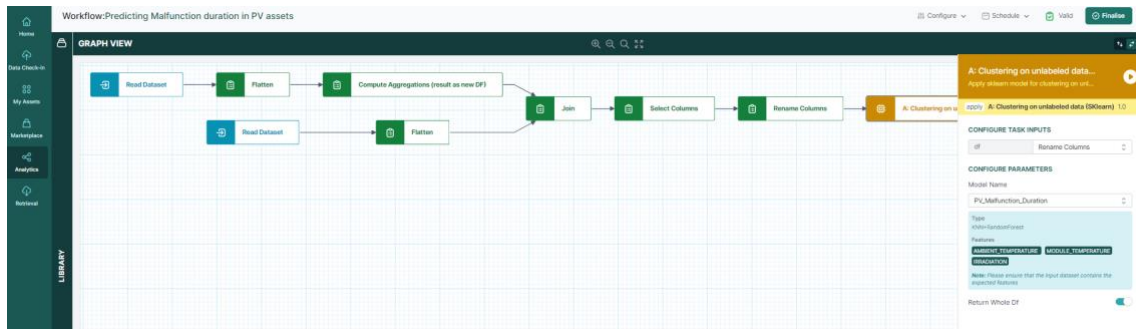


Figure 127: Model application workflow (partial view)

8.2.11.3.4 License & Availability in the SYNERGY Platform

The trained model is available in SYNERGY platform under the name of “PV_Malfunction_Duration”. The users can acquire it through the SYNERGY marketplace and integrate the trained model into their workflow through the “Clustering on unlabelled data (SKlearn)” block, under the terms of the established contract between the data asset provider (KBZ) and the data asset consumer.

The model is provided with a confidential license by its provider, KBZ, that is the owner of the model’s intellectual property rights.

8.2.11.3.5 Assumptions and Limitations

The malfunction duration prediction model is created based on weather parameters like temperature and irradiation. Since weather parameters are strongly affected by the season, using the trained model for a different period of the year, could affect the performance of the model. The registered model was trained on a specific period of the year, namely May and June. It is recommended to re-train the model, if the user wants to create a pipeline using the model, for a different season of the year.

8.2.12 Clustering of PV malfunctions/inefficiencies

Scope	Industrial Analytics
Problem	Predictive and Preventive Maintenance

Question

IND.V.3 - How can malfunctions/inefficiencies at generation asset level be grouped?

8.2.12.1 Description

Supervisory control and data acquisition (SCADA) systems are generally used in power plants to monitor the operation and performance of the installations, recording different and displaying live values. Different alerts can be configured in order to notify faults on the equipment where sensors are installed. However, there are some mistakes due to communication errors or not recognised by the system as anomaly that are not identified. Anomaly detection together with clustering methods can be used for identifying these anomalies and classifying them in order to shorten the repairing time and energy losses.

The combination of anomaly detection and clustering algorithms in energy generation aims to provide a helpful resource to identify and locate the potential faults that might not be identified during the routinary operations of power plants through a combination of unsupervised machine learning and deep learning techniques. The algorithm uses as input the energy generation and weather data of the PV power plant, which must be double normalised and returns the anomalies and cluster for each of them.

8.2.12.2 Background

Anomaly detection is one of the most known problems that have been targeted with Machine Learning, using unsupervised techniques (e.g. clustering) to identify points, events or observations that outlie from normal behaviours. One of the most effective and less computationally intensive models are the density-based algorithms, such as K-Means, Density Based Scan, K-Nearest Neighbour, etc. When used for anomaly detection, these models return the probability of each observation to be an outlier or anomaly. Combining the input and output (labels with values raging 0-1) data of anomaly detection problems, we could define as a binary classification problem.

Several models have been implemented within this field, mainly focused on anomaly detection on energy consumption and generation. An innovative model of unsupervised ML model for anomaly detection in time series using variational recurrent autoencoders with attention was proposed by Pereira in order to identify outliers in energy time series (Pereira et al., 2019). Himeur (Himeur et al., 2021) presented a review of the artificial intelligence (AI) models for anomaly detection in energy consumption in building. Moving to the energy generation side,



several research have been focused in developing algorithms for anomaly detection in power plants (Mulongo et al., 2020), and more specifically in solar PV power plants (Akiyama et al., 2015) by means of AI.

8.2.12.3 SYNERGY Implementation Details

8.2.12.3.1 *Input Data*

The datasets used to train and evaluate the model were taken from one of the COBRA's PV plants located central Spain, with almost 150,000 solar panels installed, operating for over a year. The data were pre-processed, regularising the data with different timestep, as well as applying different filters to discard empty values. The above filters resulted in obtaining a dataset of over two hundred thousand entries, with the following variables (features):

- Weather Data, including Plane of array (POA) irradiance, Ambient temperature and wind speed.
- Instantaneous energy generation at plant level in 15 minutes timestep, measured at substation.
- The peak power of the PV plant is required in order to scale the values of the energy generation.

8.2.12.3.2 *Approach*

Step 1: Feature pre-processing

- Cleaning (remove entries with missing values)
- Double normalisation of the energy generation in 15 minutes timestep, by means of the peak power installed and the solar irradiance.

Step 2: Anomaly detection and labelling

- The pre-trained ANN model was applied to the pre-processed dataset in order to identify generation anomalies in the values recorded. The results were labelled using binary classification: No anomaly (0) and anomaly (1).

Step 3: Filtering

- In order to keep just the data that were identified as anomalies, a filter was applied. The filtered dataset will be the input for the clustering algorithm on the next step.



Step 4: Anomaly classification

- With the potential anomalies identified and filtered, the problem is now framed as a classification problem to categorise the anomalies found. The developed model is an unsupervised machine learning classification model with 4 clusters (K-Means). This model was trained locally and thereafter uploaded to the SYNERGY Platform.

8.2.12.3.3 Technology

The anomaly detection of energy generation at PV plants model was developed locally. In particular, the corresponding training datasets from COBRA’s PV plant were retrieved from the plant’s SCADA and pre-processed to extract the features described in the previous sub-section and also to remove rows with missing values. The dataset was then split into train and test sets. The first was used to perform the model training and the latter was used for evaluation purposes. The model was trained locally and then uploaded to the SYNERGY platform.

An application pipeline that uses the pre-trained algorithm was created in the SYNERGY platform and is shown in Figure 128. Although the initial format of the datasets on which the model will be applied may vary, a pipeline was created to showcase how new datasets can be retrieved and pre-processed to be then fed into the trained model. The pipeline serves as guidance for how other datasets should also be processed in order for the model to be applicable. Figure 129 and Figure 130 show zoomed in snapshots from the application pipeline from the blocks that apply the anomaly detection and the anomaly clustering respectively.

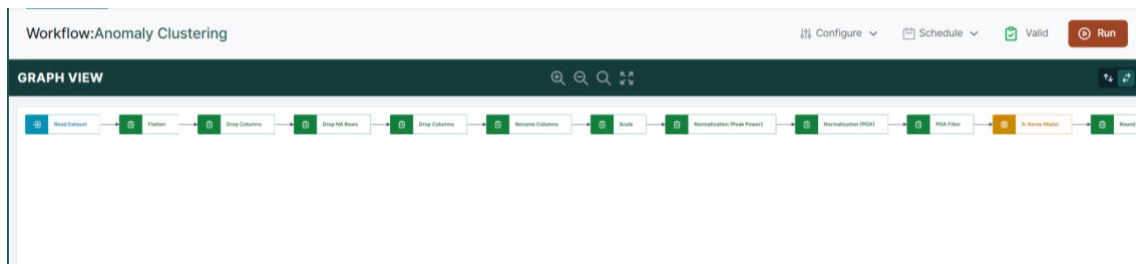


Figure 128: Application Workflow (partial view)

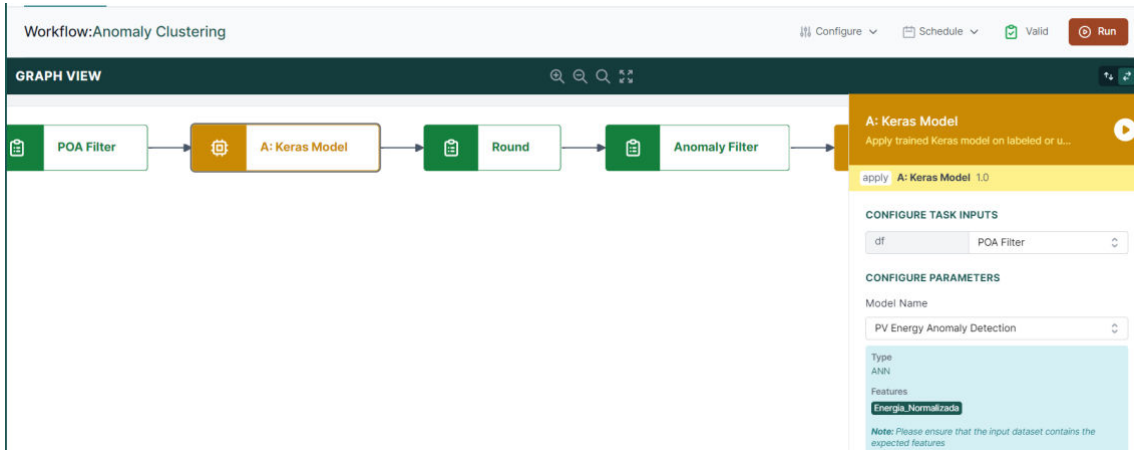


Figure 129: Application pipeline - Anomaly Detection Model

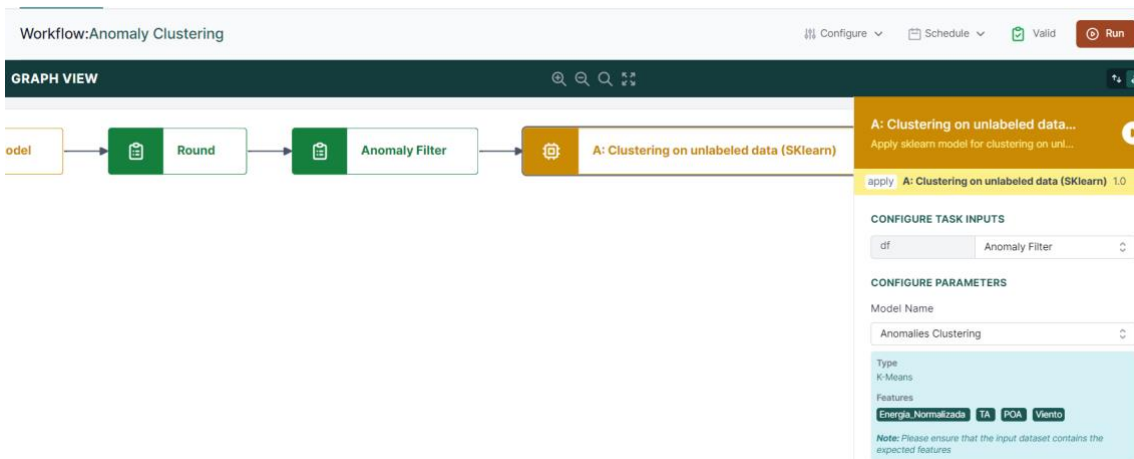


Figure 130: Application pipeline - Anomaly Clustering Model

8.2.12.3.4 License & Availability in the SYNERGY Platform

The trained model can be found in the SYNERGY platform, under the name “Anomaly Clustering”. It can be acquired through the SYNERGY marketplace and used in any pipeline.

The model is provided with a confidential license by COBRA, who is the IPR owner of the models and the pipeline.

8.2.12.3.5 Assumptions and Limitations

Due to the generic nature of the proposed solution, new features can be easily added (if available) and their contribution will be evaluated in order to improve the model performance in future releases. The trained model is available to be used within the SYNERGY platform and the description of the training dataset can help interested stakeholders understand whether their intended usage will be a good fit for the already trained model.



The number of clusters defined has been fixed in 4, one per level of devices dealing with electricity (i.e. module, combiner box, inverter, power stations) found at PV plants. This is a very generic approach, though the configuration of each PV plant may change. Nonetheless, the number of clusters may vary for each PV plant, depending on the configuration, type of modules, technology used, inverters, etc.

Another limitation is the meaning or type of anomaly that each cluster is referred to. This is something that the PV plant operator and analytics user may need to identify for each plant.



9 Conclusions & Next Steps

In this deliverable, D3.6 “SYNERGY Integrated Platform – Release 1.00”, a thorough description of the different functionalities offered by the SYNERGY Platform, is provided. In particular, the different workflows of the core platform functionalities (i.e., data check-in, data search and acquisition, data analytics) and the additional platform functionalities (CIM manager, edit organisation profile, edit user profile, wallet management) that are supported by the SYNERGY Platform, are described thoroughly as user journeys for data asset providers and consumers.

Towards this end, this deliverable documented the first official release of the SYNERGY Platform, including an overall description of the SYNERGY Platform described in Section 2, as well as all the supported functionalities that were developed under WP3 and WP4 towards the delivery of this release. In particular, the main user journeys of the developed functionalities include:

- The Data Check-in User Journey which describes thoroughly the steps from the data check-in job creation to its configuration and execution (including data harvester, mapper, cleaner, anonymiser, and encryption), and additionally the profiling of the resulting data assets.
- The Data Search and Acquisition User Journey which describes the functionalities that allow users (both data asset providers and consumers) to browse data assets (i.e. datasets, trained models and analytics pipelines results) within the SYNERGY Marketplace, and prepare acquisition contracts between data asset providers and consumers in order to retrieve such assets from the SYNERGY Platform, according to their preferences.
- The Data Analytics User Journey which describes the functionalities for designing, configuring, executing data analytics pipelines, such that the platform’s users gain valuable insights for their own and acquired data (by visualizing the appropriate results or retrieving them according to their terms). The data asset consumers are also able to register models that have been trained outside the SYNERGY Platform while they may get approval for publishing derivative assets (i.e. trained models, results) in the SYNERGY Marketplace through derivation contracts signed with the original data asset providers.
- The additional functionalities available in the SYNERGY Platform (i.e., CIM manager, the organisation profiles, the user profiles, and the wallet manager).



The deployment of the different components and services included in the current release followed a solid integration plan to ensure that their interrelations and dependencies are properly reflected and their integration is appropriately prioritised in order to provide the most added value possible to the electricity data value chain stakeholders through quick-wins while performing the initial, manual stress testing activities. The components and services that fall under the context of the current release were deployed in the SYNERGY Cloud Platform and its Secure Experimentation Playgrounds (SEP), and the Server and Edge On-Premise Environments (OPE), as described in detail in D2.7 “SYNERGY Framework Architecture including functional, technical and communication specifications v2”. Finally, the secure transfer of data across the platform’s different layers (i.e., Cloud, SEP, OPE) is deployed as mentioned in the previous integration plan (defined in D3.4 “SYNERGY Integrated Platform – Beta Release”). The integration plan (as defined in D3.4 “SYNERGY Integrated Platform – Beta Release”) remains valid, driving a major back-end development release (to be reported on M33 in D3.7 “Data Collection, Security, Storage, Governance & Management Services Bundles – Release 2.00”, D4.4 “SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles – Release 2.00”), and the final integrated platform release (to be reported on M36 with D3.8 “SYNERGY Integrated Platform & Open APIs – Release 2.00”).

In addition, this deliverable has introduced updates to the SYNERGY baseline analytics whose draft release was documented in the SYNERGY Deliverable D4.2, providing solutions to additional problems related to: Demand Forecasting, Generation Forecasting, Occupants’ Behaviour and Comfort Profiling, Flexibility Forecasting, and Predictive and Preventive Maintenance, that have been prioritised.

The current version of the SYNERGY Integrated Platform has been made gradually available to SYNERGY stakeholders taking into account the early assessment and feedback that has been collected during the development activities of the various SYNERGY energy apps in WP5-WP7, and the demonstration activities in WP8. Access to external stakeholders is also planned in coordination with the living lab activities in WP9. Further enhancements and feedback will be taken into consideration and potentially introduced in minor releases or the major upcoming versions of the platform on M36.



Annex I: References

SYNERGY Consortium. (2020). Description of Action

SYNERGY Consortium. (2021). SYNERGY D2.2 "End-user and Business requirements analysis for big data-driven innovative energy services and ecosystems v2"

SYNERGY Consortium. (2021). SYNERGY D2.7 "SYNERGY Framework Architecture including functional, technical and communication specifications v2"

SYNERGY Consortium. (2021). SYNERGY D3.1 "SYNERGY Common Information Model"

SYNERGY Consortium. (2021). SYNERGY D3.2 "Data Collection, Security, Storage, Governance & Management Services Bundles - Beta Release"

SYNERGY Consortium. (2021). SYNERGY D3.3 "SYNERGY Integrated Platform - Alpha, Mock-ups Release".

SYNERGY Consortium. (2021). SYNERGY D3.4 "SYNERGY Integrated Platform - Beta Release".

SYNERGY Consortium. (2021). SYNERGY D3.5 "Data Collection, Security, Storage, Governance & Management Services Bundles - Release 1.00"

SYNERGY Consortium. (2021). SYNERGY D4.1 "SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles - Beta Release"

SYNERGY Consortium. (2021). SYNERGY D4.2 "SYNERGY Baseline Data Analytics - Draft Release"

SYNERGY Consortium. (2021). SYNERGY D4.3 "SYNERGY Data Analytics, Sharing & Matchmaking Services Bundles - Release 1.00"



Annex II: Stress Activities Outline

Once the SYNERGY Platform Release 1.00 was deployed in the production environment, a number of stress tests were performed manually in order to check its behavior under extreme load. Such tests intended to identify the breaking point and robustness of the platform. It needs to be noted that due to the platform nature and the diverse functionalities to be tested, the use of automated testing tools was not considered as possible or relevant in this phase.

The following table presents a brief summary of the results of the stress testing activities in the current infrastructures⁶, focusing on the execution of different jobs/pipelines.

Functionality	Breaking Point
Data Check-in	<p>Batch Data Upload through Browser: File Size 1.4GB (Safari) & 1.2GB (Chrome, Firefox, Edge) due to browser limitations – Concurrent File Data Check-in Jobs: 10-40 (depending on file size).</p> <p>Batch Data Upload through the On-Premise Environments: File Size 3.5GB (peak RAM – 14GB) for a typical machine running on 8 CPUs - 16GB RAM (note: higher volumes can be uploaded through high-grade dedicated servers) – No limit for concurrent Data Check-in Jobs with on-premise storage. Concurrent Data Check-in Jobs with cloud storage: 10-40 (depending on file size).</p> <p>(3rd-party and Platform) API Data Upload: ~1800 Data Check-in Jobs per 15' (typical schedule set by the demonstrators).</p> <p>Streaming Data Upload (including execution of all data check-in steps) through Platform and 3rd-party PubSub mechanisms: ~2400 real-time messages per minute</p>
Marketplace	<p>10k Data Assets (irrespectively of whether they are datasets, models or results) in the Marketplace</p> <p>10-15 Contracts signed in the Blockchain per second</p>
Data Analytics	<p>10 concurrent (ML/DL training) workflows in Python, each running on over 5 million input records (peak RAM: 12 GB)</p> <p>40 concurrent (ML/DL application) workflows in Python, with ~100 input records</p> <p>5-10 concurrent (ML/DL) workflows in Spark, each running on ~50 million input records</p>
Data Retrieval	<p>8-10k requests in different retrieval queries per 15' (rate limits applicable per IP)</p>

⁶ One cluster with two node pools in the Google Cloud Platform: (1) 3 nodes with 2 CPUs – 8GB RAM; (2) 3-10 nodes with 4 CPUs – 16GB RAM. Depending on the workload by the demonstrators and external stakeholders, the infrastructures will obviously scale accordingly.

