Toward an Energy Data Platform Design: Challenges and Perspectives from the SYNERGY Big Data Platform and AI Analytics Marketplace

Fenareti Lampathaki, Evmorfia Biliri, Tasos Tsitsanis, Kostas Tsatsakis, Dimitris Miltiadou, and Konstantinos Perakis

Abstract Today, the need for "end-to-end" coordination between the electricity sector stakeholders, not only in business terms but also in securely exchanging real-time data, is becoming a necessity to increase electricity networks' stability and resilience while satisfying individual operational optimization objectives and business case targets of all stakeholders. To this end, the SYNERGY energy data platform builds on state-of-the-art data management, sharing, and analytics technologies, driven by the actual needs of the electricity data value chain. This paper will describe the layered SYNERGY Reference Architecture that consists of a Cloud Infrastructure, On-Premise Environments, and Energy Apps and discuss the main challenges and solutions adopted for (a) the design of custom pipelines for batch and streaming data collection and for data manipulation and analytics (based on baseline or pre-trained machine learning and deep learning algorithms) and (b) their scheduled, on-event, or real-time execution on the cloud, on-premise and in gateways, toward an energy data space. Particular focus will be laid on the design of the SYNERGY AI analytics marketplace that allows for trustful sharing of data assets (i.e., datasets, pipelines, trained AI models, analytics results) which belong to different stakeholders, through a multi-party smart contract mechanism powered by blockchain technologies.

Keywords Energy Data Spaces · Energy data platform · Architecture · Data pipelines · Data lifecycle management · Data sharing · Data analytics

F. Lampathaki (✉) · E. Biliri · T. Tsitsanis · K. Tsatsakis
Suite5 Data Intelligence Solutions, Limassol, Cyprus
e-mail: fenareti@suite5.eu

D. Miltiadou · K. Perakis
UBITECH, Thessalias 8 & Etolias, Chalandri, Greece

1 Introduction

The European electricity sector is undergoing a major fundamental change with the increasing digitalization and roll-out of smart meters. This advent of the electricity sector modernization comes together with the fact that the power system is becoming more thoroughly monitored and controlled from "end to end" and through the whole value chain of stakeholders involved in the electricity system operation. This is a huge shift away from traditional monitoring and control approaches that have been applied exclusively over the transmission and distribution networks, since the smart electricity grid era is pushing sensing, control, and data collection at the edge of electricity networks, which needs to be further re-defined due to the wide penetration of distributed energy resources (DERs), such as renewable energy sources (RES), smart home devices and appliances (IoT-enabled), distributed storage, smart meters, and electric vehicles (EVs).

Distributed smart grid resources are associated with the generation of vast amounts of data, spanning SCADA systems information (generation, transmission, and distribution), smart metering and sub-metering information (demand), IoT device information, distributed generation and storage data, electric vehicle, and electricity market information, altogether characterized by continuously increasing growth rate, multi-diverse spatio-temporal resolutions, and huge volume. Such large datasets provide significant opportunities for better "end-to-end" monitoring, control, and operation of electric grids by enabling better understanding and offering further insights on all aspects affecting (directly or indirectly) the operation of the networks (and DERs, as new individual connected components of smart electricity grids) toward optimizing their performance (individually and network-wide), through advanced big energy data analytics [1, 2]. However, while the industry may now recognize the potential of Big Data, it struggles to translate that into action. A recent study from CapGemini [1] found that only 20% of smart grid stakeholders have already implemented Big Data analytics. There is a significant group (41%) with no Big Data analytics initiatives which compares quite unfavorably with take-up levels in other sectors.

The analytics opportunity for electricity sector stakeholders is there, and benefits are significant; however, recent studies have pointed out that electricity sector actors are reluctant to make the move due to high upfront costs and sheer complexity of data [1]. Taking data management and analytics away from their hands (in a trustful manner, thus reducing complexity and changing their mindsets) and offering to them easily digestible intelligence extracted from the advanced processing and analysis of highly diverse, variable, and volatile data streams (through ready-to-use trained algorithms that can be easily utilized in different contexts and business cases) could be the first step forward, toward enabling the realization of data-driven optimization functions that can pave a ROI-positive path to effectively solving operational and business challenges and highlighting the value of the big distributed data generated at the wealth of end points of the power system.

The value of similar approaches and implementation references has been already showcased in relevant reference implementations, mainly in the USA, where the introduction of advanced (and near-real-time) data analytics in the electricity sector proved to facilitate the early detection of anomalies, trends, possible security breaches, and other costly business interruptions and enable the avoidance of undesired costs, along with the creation of new profit opportunities [3]. The cost, investment, and resulting value of Big Data analytics and data-driven insights have different impacts on the grid's various major stakeholders. Each stakeholder has different roles and responsibilities in managing, storing, processing, protecting, owning, and using data. For instance, the value of Data Spaces and analytics for electricity grid operators lies on the fact that they can further optimize the operational stability and resilience of their network through improved demand and generation forecasting, advanced predictive maintenance, and management of their owned assets (lines, transformers, sub-station equipment, etc.), improve power quality and continuity of supply by avoiding interruptions due to equipment failures, optimize scheduling of maintenance activities, and enhance physical security of critical distribution network infrastructure.

In this context, this paper introduces the SYNERGY Reference Architecture that aims to allow electricity value chain stakeholders to simultaneously enhance their data reach and improve their internal intelligence on electricity-related optimization functions while getting involved in novel sharing/trading models of data sources and intelligence, in order to gain better insights and shift individual decision-making at a collective intelligence level. The SYNERGY Reference Architecture is based on state-of-the-art approaches from a technology perspective (in terms of data management, data analytics, data sharing, and data security techniques and technologies), as well as from a market perspective (considering the different data platforms that are introduced in Sect. 2). The different workflows that are enabled though the SYNERGY Reference Architecture are discussed highlighting the core challenges that have been jointly identified by representatives of the electricity data value chain and technology experts.

This chapter relates to the technical priorities of the European Big Data Value Strategic Research and Innovation Agenda [4], addressing the horizontal concerns "Data Management," "Data Protection," "Data Processing Architectures," "Data Analytics," and "Data Visualization" and the vertical concern "Industrial Data Sharing Platforms" of the BDV Technical Reference Model. In addition, the chapter relates to the "Knowledge and Learning" and "Reasoning and Decision Making" enablers of the AI, Data and Robotics Strategic Research, Innovation and Deployment Agenda [5].

2 Data Platforms

The unprecedented supply of data and the technological advancements in terms of storage and processing solutions, e.g., offered through on-demand computing

power as a service through the cloud, are among the forces fueling the emergence of data as a new tradable good online. Data marketplaces and Data Spaces are the infrastructures through which this new market is realized. The market's growth however cannot be attributed solely to the technical innovations, notwithstanding their enabling role, but should be examined under the prism of demand and supply. The abundance of data created every day and the way data analysis can transform them to insights for more informed decision-making create incentives for businesses to develop a data sharing mentality and investigate data monetization approaches. Technical, motivational, economic, political, legal, and ethical challenges in fostering a data sharing mentality in an industry environment are numerous, yet realizing the prospective benefits from disrupting the current data siloed situation is an important first step toward seeking ways to overcome the aforementioned barriers.

A more concrete definition of a marketplace would be that of a "multi-sided platform, where a digital intermediary connects data providers, data purchasers, and other complementary technology providers" [6]. In practice, functionalities of data marketplaces extend beyond the implementation of the data trading action.

2.1 Generic-Purpose Data Hubs and Marketplaces

A number of leading data marketplaces have emerged over the last years, demonstrating significant diversity in the provided offerings, stemming from the target domain and scope and the underlying technologies. The data marketplace concept is inherently interdisciplinary, in the sense that it brings together technological, legal, and business knowledge in order to successfully capture and satisfy the underlying demand and supply data needs. In many cases, the marketplace services constitute an application of an underlying technology, built to support the data trading functionalities, but also independently exploited.

Indicative examples of data marketplaces are briefly presented below in order to give a comprehensive overview of the current status and future perspectives of these platforms and outline ways in which they could create an innovation environment for new digital business models:

- Datapace (https://datapace.io/) is a marketplace for IoT sensor data with technical and policy-based data verification and access to a worldwide network of sensors. It supports micropayments using a custom token (namely, the TAS which is native to the platform and has no use externally to it) and offers smart contracts based on a permissioned enterprise blockchain. The Datapace storage encrypts and anonymizes the access to the submitted data streams.
- The DX Network (https://dx.network/) is one of the largest blockchain-based business data marketplaces. It is API-based, therefore can be easily integrated into any data-enabled services, and focuses on real-time data streams, allowing asset trading at data point granularity which is based on its custom query language that leverages Semantic Web technologies.

- Dawex (https://www.dawex.com/en/) is a leading data exchange technology company and the operator of one of the largest global data marketplaces. Its global marketplace provides customizable data access control mechanisms, supports various data formats, and provides visualizations to evaluate data quality and contents. Representative data samples are created through custom algorithms to support this process. Data are hosted encrypted, and the platform has certification from independent data protection authorities to ensure regulatory compliance. Dawex also enables organizations to create their own data exchange platforms using its technology. Apart from the core data trading services, the platform offers machine learning algorithms to match data supply and demand, allowing for proactive suggestions to members.
- IOTA (https://data.iota.org/#/) is an open, feeless, and scalable distributed ledger, designed to support frictionless data and value transfer. IOTA's network, called Tangle, immutably records exchanges and ensures that the information is trustworthy and cannot be tampered with or destroyed and was designed to address blockchain inefficiencies in terms of transaction times and scalability. It is a secure data communication protocol and zero-fee microtransaction system for the IoT/M2M.
- Qlik DataMarket (https://www.qlik.com/us/products/qlik-data-market) offers an extensive collection of up-to-date and ready-to-use data from external sources accessible directly from within the company's data analytics platform Qlik Sense. It provides current and historical weather and demographic data, currency exchange rates, as well as business, economic, and societal data, addressing data augmentation needs in the contextualization and analysis of business data leveraging external sources. Integration is in this context effortless, and validation, profiling, and quality measures are provided to evaluate the data available in the market.
- Streamr (https://streamr.network/marketplace) offers a marketplace for real-time data, leveraging blockchain and Ethereum-based smart contracts for security-critical operations like data transfers. It provides tools and libraries to (a) create, process, visualize, and sell real-time data and (b) acquire and ingest real-time data to enable business intelligence. The marketplace is an application of the Streamr network, a massively scalable peer-to-peer network for transporting machine data in real time with the PubSub pattern. It also offers crowdsourcing functionalities to incentivize gathering of previously unavailable data.
- MADANA (https://www.madana.io/vision.html) aims to create a self-governing and community-driven market for data analysis through a platform that connects data providers, data analysis providers (called plugin providers in the platform's terminology), and consumers/buyers for data analysis results. Beyond a marketplace, MADANA aspires to be a platform for data analysis which provides secured computation, data monetization, and the outsourcing of analytics on demand. Purchases are based on smart contracts and the platform's custom cryptocurrency called MADANA PAX. Upon collection, data are encrypted and kept in a distributed storage. Access is not foreseen to be provided to raw data, so only analysis results can be purchased.

European projects have been also active in this promising field. The ICARUS (www.icarus2020.aero) [7] marketplace offers brokerage functionalities specialized in aviation data assets conforming to a common data and metadata model and provides smart contracts based on Ethereum. Safe-DEED (https://safe-deed.eu/) explores how technology, e.g., in the fields of cryptography and data science, can foster a data sharing mentality, incentivizing businesses and innovating business models.

Distributed ledger technology (DLT) applications are also extremely popular, showcasing numerous additional data marketplaces, e.g., (a) Wibson (https://wibson.org/), a decentralized blockchain-based marketplace allowing members of the public to profit from securely and anonymously selling their personal data, and (b) Datum (https://datum.org/), which enables decentralized storage of structured data on a smart contract blockchain and data brokerage using a smart token.

Depending on the type of data, both in terms of content and formats, the prospective buyers and sellers, the target industries, the employed technologies, etc., a long list of platforms offering data marketplace services, either exclusively or as a side product of their core/other businesses, can be compiled. When traded commodities extend beyond data to other data-based assets, e.g., processed data and extracted insights, the number of platforms that can be considered as relevant can easily explode. Identifying and examining all data marketplaces is not possible and would largely be out of scope for the current work. However, different approaches used in literature to study and group data marketplaces have been extensively studied in [8–11].

For many-to-many data marketplaces, additional attributes could be selected for a more fine-grained analysis, e.g., the choice between a centralized and a decentralized design. This architecture decision entails implementation implications and affects the overall marketplace operation in various ways. Indicatively, [11] highlight that in the centralized setting, the market intermediary trades off quality (provenance control) for lower transaction costs. In a decentralized setting, e.g., one implemented through a distributed ledger technology, transaction costs are higher and bottlenecks may emerge, yet there is increased provenance control and transparency.

An important attribute of data marketplaces is the contract drafting and enforcement process, which is typically one of the services provided by such platforms and is an integral part of asset trading. Stringent enforcement of contract terms in this scope is challenging, and several factors, including technical limitations and legal implications, need to be examined. Data protection and security mechanisms, as well as data privacy and confidentiality, should be ensured to foster trust among the platform members and to comply with applicable regulations, e.g., the General Data Protection Regulation (GDPR). Technical advancements can also help in this direction, e.g., multi-party computation (MPC), a cryptographic technique that enables joint data analyses by multiple parties while retaining data secrecy, is explored as a way to increase industry's willingness to participate in data marketplaces. Auditability should also be possible in industry data trading agreements, yet anonymity in transactions may also be required. Furthermore, licensing, ownership, and IPR

of data and data products are contentious issues requiring careful definition and adjudication, which may not be possible to capture within blanket agreements [11]. License compatibility, in the case of combination and derivation, e.g., data assets as a result of data integration from multiple sources and/or data analysis processes, is also challenging. On a final note, for a marketplace to establish a vibrant business ecosystem that will render it sustainable, data interoperability achieved through agreed data and metadata models and common semantics is required. Especially in the case of data marketplaces connecting numerous suppliers and consumers, data discoverability, timely and secure exchange, effortless ingestion, and (re-)usability across diverse data sources, all facilitated by an appropriate level of automation, will allow the marketplace to scale and foster opportunities on monetizing data. Such considerations were taken into consideration in the scope of the SYNERGY positioning.

2.2 Energy Data Hubs and Marketplaces

In the energy domain, there is no mature state of the art about the role and potential data marketplaces. The recently evolving energy data hubs, though, provide some insights from research initiatives about the potential of the energy data marketplaces in the future.

The term energy data hub, or energy data space, is defined as an on-demand, back-end repository of historical and current energy data. The objective is to streamline energy data flows across the sector and enable consumers, authorized agents on consumer's behalf, and other users to access energy data. While there is an increasing interest about the penetration of energy data hubs, following the increased installation of smart equipment and the deregulation of the market in the energy value chain, the number of the existing implementations is rather narrow. The data hubs are mainly focusing on specific business stakeholders and business processes in the energy value chain [12], and thus a business-driven taxonomy of the different energy hubs is considered as follows:

- *Retail data/smart meter hubs* are defined as the data hubs at EU country level which are responsible for the management of smart metering data. Retail data hubs are introduced to address two primary issues: (a) secure equal access to data from smart metering and (b) increase efficiency in the communication between market parties, especially between network operators and retails for billing and switching purposes. There are many region-level implementations around the world considering the smart meter's deployment with the most prominent examples being:

  - The Central Market System (CMS) aka ATRIAS started in 2018, as the centralized data hub to facilitate the data exchange between market parties in Belgium. The CMS focuses on the data exchange between the DSOs and retail businesses and thus connects the databases of the network operators (who

collect the data from the smart meters) with the relevant and eligible market parties. Other parties, like the transmission system operators and third-party service providers, may access the data as well.

– In Norway [13], the ElHub (Electricity Hub) facilitates the data exchange between market parties in Norway. ElHub is operated by the national TSO with the smart metering data to be collected via the DSOs and stored in the ElHub together with consumer data from the retailers. The customers are in full control of their data, which they can access via an online tool and thereby manage third-party access to their datasets.

- *Smart market data hubs* are defined as the data hubs at EU country level responsible for the management of energy market data. The major electricity market operators in Europe are handling energy market data hubs to share data with the different business stakeholders. Special reference can be made to the following market operators:

  – Nord Pool (https://www.nordpoolgroup.com/services/power-market-data-services/) which runs the leading power market in Europe and offers day-ahead and intraday markets to its customers. The day-ahead market is the main arena for trading power, and the intraday market supplements the day-ahead market and helps secure balance between supply and demand. Access on real-time market data is available online, though fine-grained data services (access on data per country, product, means of access, etc.) are offered by the company. More specifically, customized power data services may be provided to external interest parties, setting that way a market-based framework for data exchange.
  – EPEX SPOT energy market data hub (https://www.epexspot.com/en) which offers a wide range of datasets covering the different market areas, available through different modalities: from running subscriptions for files that are updated daily to access to one-shot historical data.

- *Smart grid data hubs*: This is a step beyond the currently deployed smart meter data hubs. Through their evolving role around Europe, the network operators aim to act as data hub providers beyond smart meter data, while their data hubs will be used to provide services for the network operators (e.g., data exchange between the DSO and the TSO) as well as for new market entrants with new business models (e.g., related to behind-the-meter services). Therefore, the role of network operators as grid-related data managers is expanding. Under this category, there are some very promising initiatives, which are further presented below:

  – At country/regional level, there are state network operators responsible to publish their data required for the normal operation of the grid. Toward this direction, the European Network of Transmission System Operators for Electricity (ENTSOE) is operating a Transparency Platform (https://transparency.entsoe.eu/) where the data from the national TSOs are published in order to facilitate the normal operation of the transmission grid in Europe.

  – At the regional level, the distribution network operators have started making their data public to help other stakeholders and market parties with, e.g., better decision-making, create new services, and promote synergies between different sectors. As not all DSO data are suitable to be made publicly available due to potential breaches of security or violations of privacy regulations, it is important for DSOs to have a common understanding. For that reason, E.DSO made recently available a policy brief to illustrate the possibilities of open data from each member state, in terms of meaningful use cases [14]. Key highlights of open data repositories from DSOs (EDP in Portugal, ENEDIS in France) are to be considered for the future expansion of open data repositories in the EU.
  – Moving beyond the national level is the PCI project of Data Bridge (now defined as an Alliance of Grid Operators, https://www.databridge.energy/) with the goal to ensure the interoperability of exchanging different types of data between a variety of stakeholders (like system operators, market operators, flexibility providers, suppliers, ESCOs, end customers). Types of data may include smart meter data (both low-voltage and high-voltage meter data), sub-meter data, operational data, market data required for functioning flexible energy market, reliable system operation, etc.

From the aforementioned analysis, it is evident that the main focus of the energy actors in the data management landscape is about establishing functional energy data hubs that will be able to provide useful information to selected stakeholders of the energy value chain in a unified way. The concept of enabling the different energy stakeholders to match and trade their energy data assets and requirements in a marketplace environment does not exist yet at large scale. There are some early implementations of generic data marketplaces that enable management of data from the energy sector, which include (in addition to Dawex that has been already analyzed in Sect. 2.1 and includes an energy-specific solution with focus on smart home and renewable source data):

• Snowflake data marketplace (https://www.snowflake.com/datasets/yes-energy) is a data hub that enables data providers to leverage and monetize their data. In this platform, Yes Energy, the industry leader in North American power market data and analytic tools, acts as a data provider in the platform by collecting, managing, and continuously delivering real-time and historical power market data series including market data, transmission and generation outages, real-time generation and flow data, and load and weather forecasts.
• The re.alto marketplace (https://realto.io/) represents the first mature attempt to provide a European API marketplace for the digital exchange of energy data and services. Established in 2019, re.alto data marketplace enables companies to capture, organize, and share data and services easily, quickly, and securely. So far, the datasets available in the platform span between energy market data, asset generation data, weather data, energy metering, and smart home data. In addition, AI applications such as generation and demand forecasts, price forecasts, etc. are made available through the marketplace.

- The ElectriCChain (http://www.electricchain.org) is defined as an Open Solar data marketplace with an initial focus on verifying and publishing energy generation data from the ten million solar energy generators globally on an open blockchain. The ElectriCChain project supports the development of open standards and tools to enable generation asset owners to publish solar electricity generation data and scientists, researchers, and consumers to have access on the data and insights they need.

On the other hand, in the field of IoT solutions (as the wider term that covers the smart assets deployed in the electricity network spanning from network devices, smart meters, home automation solutions, DER data loggers, etc.), there is an ongoing discussion about the importance of the data and the way to put IoT data to work and cash, offering the information to third parties through data marketplaces. There are many small-scale/proof-of-concept initiatives of IoT data marketplaces to collect sensor data which data providers source from smart home appliances and installations in people's homes and smart cities, while companies looking to understand consumer behavior can leverage such machine data directly from the marketplaces in real time. The most prominent solutions include Datapace (https://datapace.io/) that offers blockchain-powered secure transactions and automated smart contracts to sell and buy data streams from any source, physical assets, autonomous cars, drones, and the IOTA marketplace that has been also mentioned in Sect. 2.1.

From the aforementioned analysis, it is evident that the concept of regulated and standardized energy data marketplaces is new for a domain that is still undergoing its digital transformation. There is an ongoing work to design and develop standards-based data hubs to ensure interoperability of exchanging different types of data between a variety of energy stakeholders, but still the value of such data that can be made available via data platforms and marketplaces remains largely unexplored.

3 SYNERGY Reference Architecture

In an effort to leverage such unique data-driven opportunities that the electricity data value chain presents, our work is focused on the development of an all-around data platform that builds on state-of-the-art technologies, is driven by the actual needs of the different stakeholders, and turns over a new leaf in the way data sharing and data analytics are applied. Taking into consideration the different use cases and requirements of the different energy stakeholders as well as the state of play described in Sect. 2, the reference architecture of the overall SYNERGY platform has been conceptually divided into three main layers as depicted in Fig. 1:

- The **SYNERGY Cloud Infrastructure** that consists of (a) the *Core Big Data Management Platform*, essentially including the Energy Big Data Platform and the AI Analytics Marketplace which are instrumental for all functionalities that SYNERGY supports at all layers, and (b) the *Secure Experimentation*
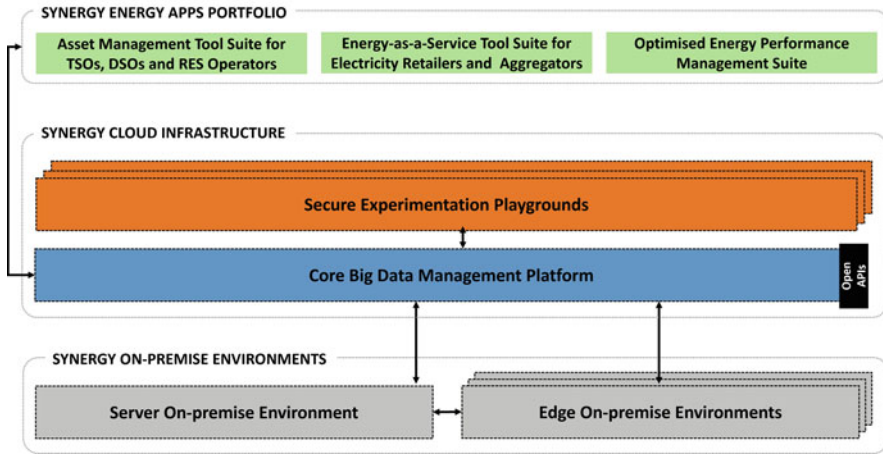
Fig. 1 SYNERGY three-layered high-level architecture

Playgrounds (SEP) which are realized in the form of dedicated virtual machines that are spawned per organization to ensure that each electricity data value chain stakeholder is able to execute Big Data analytics in isolated and secure environments in the SYNERGY Cloud Infrastructure.

- The **SYNERGY On-Premise Environments (OPE)** which are executed in the energy stakeholders' premises for increased security and trust and can be distinguished in the server environment and the edge environments that are installed in gateways. The On-Premise Environments are not self-standing, but always communicate with the SYNERGY Cloud Infrastructure to deliver their intended functionality.
- The **SYNERGY Energy Apps Portfolio** that embraces the set of applications addressed to the needs of (a) DSOs (distribution system operators), TSOs (transmission system operators), and RES (renewable energy sources) operators in respect to grid-level analytics for optimized network and asset management services, (b) electricity retailers and aggregators for portfolio-level analytics toward energy-as-a-service (EaaS) solutions, and (c) facility managers and ESCOs (energy service companies) toward building/district-level analytics from the perspective of optimized energy performance management.

In order to deliver the intended functionalities toward the different electricity data value chain stakeholders who at any moment may assume the role of data asset providers and/or data asset consumers, the high-level architecture consists of the following data-driven services bundles that have well-defined interfaces to ensure their seamless integration and operation within the SYNERGY integrated platform:

- **Data Collection Services Bundle** which enables the configuration of the data check-in process by the data provider at "design" time in the Core Big Data Management Platform and its proper execution in the SYNERGY Cloud Infrastructure and/or the On-Premise Environments. Different data ingestion, mapping, and transformation and cleaning services are invoked to appropriately handle batch, near-real-time, and streaming data collection.
- **Data Security Services Bundle** that is responsible for safeguarding the data assets in the overall SYNERGY platform (i.e., Core Big Data Management Platform and On-Premise Environments for end-to-end security) through different ways, e.g., by anonymizing the sensitive data (from an individual or business perspective), by selectively encrypting the data, and by applying access policies over the data assets that allow a data provider to control who can even view them.
- **Data Sharing Services Bundle**, essentially providing the SYNERGY Core Big Data Management Platform with the functionalities expected from a data and AI analytics marketplace in terms of sharing and trading data assets (embracing datasets, pre-trained AI models, analytics results) in a secure and trustful manner, powered by the immutability and non-repudiation aspects that are available in distributed ledger technologies.
- **Data Matchmaking Services Bundle** that delivers exploration and search functionalities (in the SYNERGY Core Big Data Management Platform) over data assets that the data consumers are eligible to view and potentially acquire while providing recommendations for additional data assets of interest or for electricity data value chain stakeholders who could potentially have/create the requested data asset.
- **Data Analytics Services Bundle** which lies at the core of the design of data analytics pipelines including the data manipulation configuration, the basic and baseline (pre-trained) machine learning and deep learning algorithms configuration, and the visualization/results configuration, in the SYNERGY Core Big Data Management Platform, while allowing for the execution of the defined pipelines in the Secure Experimentation Playgrounds and the On-Premise Environments.
- **Data Storage Services Bundle** that offers different persistence modalities (ranging from storage of the data assets, their metadata, their indexing, the algorithms and pipelines, the contracts' ledger, etc.) depending on the scope and the type of the data in the SYNERGY Cloud Infrastructure (in the Core Cloud Platform and the Secure Experimentation Playgrounds) and the On-Premise Environments.
- **Data Governance Services Bundle** that provides different features to support the proper coordination and end-to-end management of the data across all layers of the SYNERGY platform (cloud, on-premise).
- **Platform Management Services Bundle** which is responsible for resources management, the security and authentication aspects, the notifications management, the platform analytics, and the Open APIs that the SYNERGY platform provides.
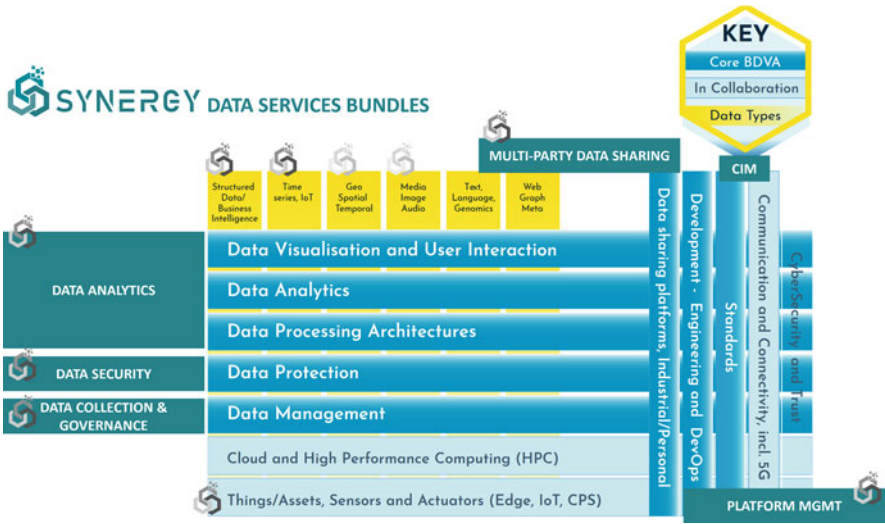
Fig. 2 SYNERGY Data Services Bundles in relation to the BDVA Reference Model

As depicted in Fig. 2, the overall SYNERGY Big Data Platform and AI Marketplace, along with its different Data Services Bundles, is well aligned to the BDVA Reference Model defined in the European Big Data Value Strategic Research and Innovation Agenda [15]. On the one hand, topics around Data Management are appropriately addressed through the SYNERGY Data Collection and Data Governance Service Bundles. Data Protection is considered from an all-around perspective in the SYNERGY Data Security Service Bundle. Data Processing Architectures, Data Analytics, and Data Visualization and User Interaction aspects have a similar context and orientation as in the SYNERGY Data Analysis Services Bundle. On the other hand, the Data Sharing Platforms are indeed tackled through the SYNERGY Data Sharing Services Bundle that is innovative in introducing the concept of multi-party sharing. Development, Engineering, and DevOps aspects are well embedded in the SYNERGY Platform Management Services Bundle. Finally, the Standards dimension is addressed within the SYNERGY Common Information Model that builds upon different energy data standards, ontologies, and vocabularies.

It needs to be noted that the SYNERGY architecture was designed taking into consideration the SGAM philosophy and design patterns [16, 17] even though in a more loosely coupled manner.

3.1 SYNERGY Cloud Infrastructure Layer

As depicted in Fig. 3, the *SYNERGY Core Big Data Management Platform* (or SYNERGY Core Cloud Platform in abbreviation) is the entry point for any user (as representative of an electricity data value chain stakeholder) in the overall SYNERGY platform. In order to check in data to the SYNERGY platform, the **Data Handling Manager** in the SYNERGY Core Cloud Platform provides the user interfaces to properly configure and manage the data check-in jobs at "design" time, according to the settings and preferences of each data provider for uploading batch data as files; collecting data via third-party applications' APIs, via open data APIs, or via the SYNERGY platform's APIs; and ingesting streaming data (through the SYNERGY platform's mechanisms or through the stakeholders' PubSub mechanisms). Upon configuring the data ingestion step, the data providers need to properly map the sample data they have uploaded to the SYNERGY Common Information Model (CIM) following the suggestions and guidelines of the **Matching Prediction Engine**. The SYNERGY Common Information Model is built on different standards, such as IEC 61968/61970/62325, IEC 61850, OpenADR2.0b, USEF, and SAREF, and aims to provide a proper representation of the knowledge of the electricity data value chain, defining in detail the concepts to
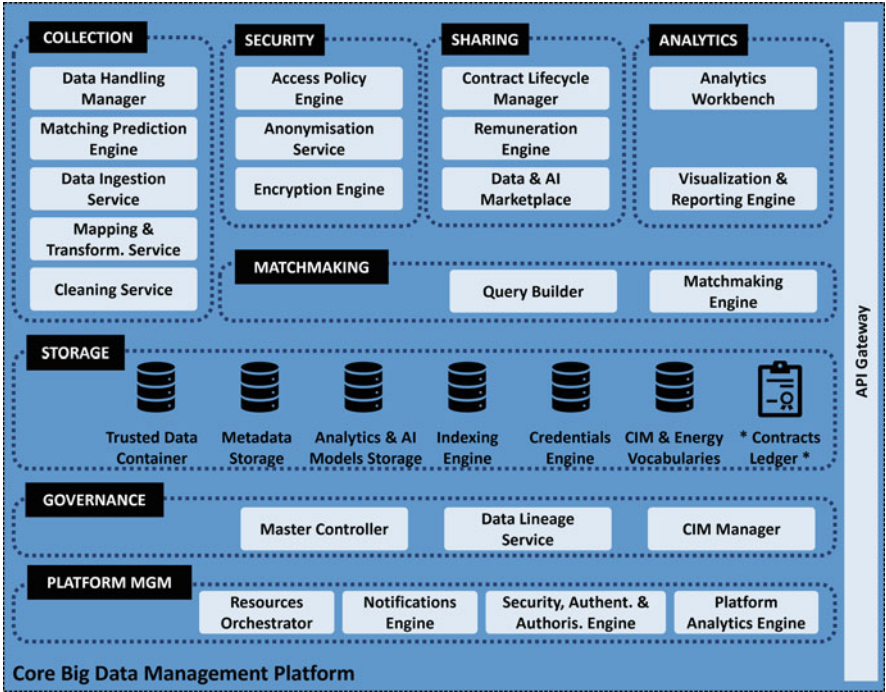

Fig. 3 Detailed component view of the SYNERGY Core Cloud Platform

which the datasets that are expected to be uploaded in the SYNERGY marketplace will refer and taking into consideration the standards' modelling approaches.

Optionally, the data providers are able to also configure the cleaning rules, the anonymization rules, and the encryption rules that need to be applied over the data. The **Access Policy Engine** provides the opportunity to define access policies based on different attributes in order to fully control which stakeholders can potentially view the specific data asset's details in the SYNERGY platform.

The data check-in job execution is triggered by the **Master Controller** according to the schedule set by the data providers and in the execution location they have set (i.e., Cloud Platform or On-Premise Environment). The Master Controller communicates with the **Resources Orchestrator** to ensure the necessary compute and memory resources (esp. in the SYNERGY Cloud Infrastructure) and orchestrates the appropriate list of services among the **Data Ingestion Service**, the **Mapping and Transformation Service**, the **Cleaning Service**, the **Anonymization Service**, and the **Encryption Engine** that are invoked in a sequential manner while forwarding them the data check-in job's configuration. The data are stored in **Trusted Data Containers** in the Data Storage Services Bundle, and a set of metadata (in alignment with the SYNERGY metadata schema built on DCMI and DCAT-AP) are either extracted automatically during the previous steps (e.g., in the case of temporal coverage, temporal granularity, spatial coverage, and spatial granularity metadata that can be extracted from the data, as well as the data schema mapped in the SYNERGY CIM) or manually defined by the data providers in the **Data and AI Marketplace** (such as title, description, tags, and license-related metadata) and persisted in the **Metadata Storage**.

The Data and AI Marketplace is essentially the one-stop shop for energy-related data assets from the electricity data value chain stakeholders as it enables secure and trusted data asset sharing and trading among them. It allows them to efficiently search for data assets of interest through the **Query Builder** and provides them with the help of the **Matchmaking Engine** with recommendations for data assets or data assets' providers (that may potentially have/create the specific data asset). The Data and AI Marketplace allows data consumers to navigate to the available data assets, preview their offerings, and proceed with their acquisition through smart data asset contracts that are created, negotiated, and signed among the involved parties in the **Contract Lifecycle Manager** and stored in each step in the **Contracts Ledger**. The contract terms of use, the cost and VAT, the contract effective date, the contract duration, the data asset provider, and the data asset consumer are among the contract's details that are stored in hash in the blockchain. In order for a signed contract to be considered as active, the respective payment needs to be settled with the help of the **Remuneration Engine**.

In order for electricity data value chain stakeholders to leverage the potential of data analytics over data that they own or have acquired, the **Analytics Workbench** gives them the opportunity to design data analysis pipelines according to their needs and requirements. Such pipelines may consist of (a) different data manipulation functions, (b) pre-trained machine learning or deep learning algorithms that have been created for the needs of the energy domain, or (c) simple algorithms that are

offered in an out-of-the box manner wrapping the Spark MLlib algorithms, the sk-learn algorithms, and the TensorFlow (over Keras) algorithms. The execution settings are defined by the data asset consumers that define when and how the data analysis pipeline should be executed and how the output will be stored. In this context, the **Visualization and Reporting Engine** allows the data asset consumers to select, customize, and save appropriate visualizations to gain insights into the analytics results, but also to create simple reports to potentially combine results.

The **API Gateway** allows the authorized SYNERGY energy applications and any application to retrieve from the SYNERGY platform's Open APIs the exact raw data or analytics results they need according to filters they are able to set. The overall platform's security, organization's and user's registration, and authorization decisions are dependent on the **Security, Authentication and Authorization Engine**.

The SYNERGY cloud platform is complemented by the **Data Lineage Service** to provide provenance-related views over the data assets; the **Notifications Engine** to send notifications about the ongoing processes that are related to a user or organization; the **Platform Analytics Engine** that provides insights into the added value of the data assets in the SYNERGY platform, but also on the overall platform's services progress; and the **CIM Manager** that is behind the evolution and propagation of changes of the Common Information Model across the involved services in the whole SYNERGY platform.

The execution of a data analysis job in the SYNERGY Cloud Platform is performed in *Secure Experimentation Playgrounds* which are essentially sandboxed environments that become available per organization. The data that belong to an organization or have been acquired by an organization (based on a legitimate data asset contract) are transferred through the **Data Ingestion Service** based on the instructions provided by the Master Controller, are decrypted upon getting access to the decryption key in the **Encryption Engine** (with the help of the Master Controller and the Security, Authentication and Authorization Engine), and are stored in **Trusted Data Containers**. Any data analysis pipeline that needs to be executed is triggered according to the organization's preferences by the Master Controller that invokes the **Data Manipulation Service** and the **Analytics Execution Service**. The **Secure Results Export Service** is responsible to prepare the results for use by the respective organization in different ways (e.g., as a file, exposing them via an API, sharing them in the Data and AI Marketplace). Finally, the **Data Lineage Service** provides an overview of the relations and provenance of the data assets stored in the Secure Experimentation Playground (as depicted in Fig. 4).

3.2 SYNERGY On-Premise Environments Layer

The *SYNERGY Server On-Premise Environment* is responsible for (a) preparing the data assets, which an organization owns, "locally" to ensure end-to-end security (especially when encryption is required in the data check-in job configuration)
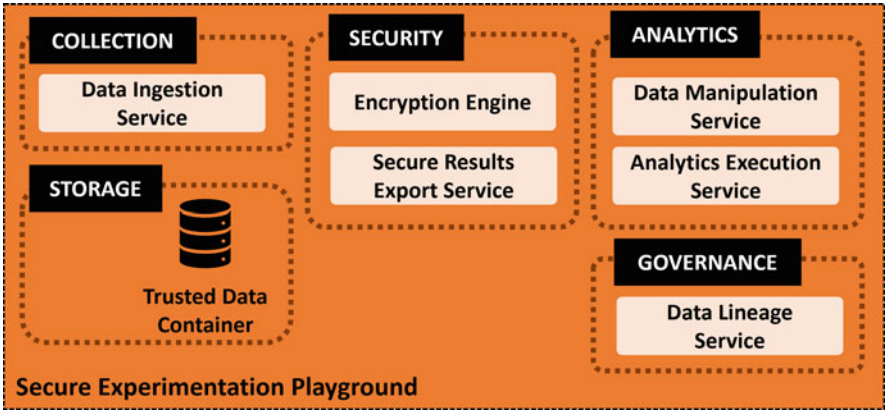
Fig. 4 Detailed component view of the SYNERGY Secure Experimentation Playground
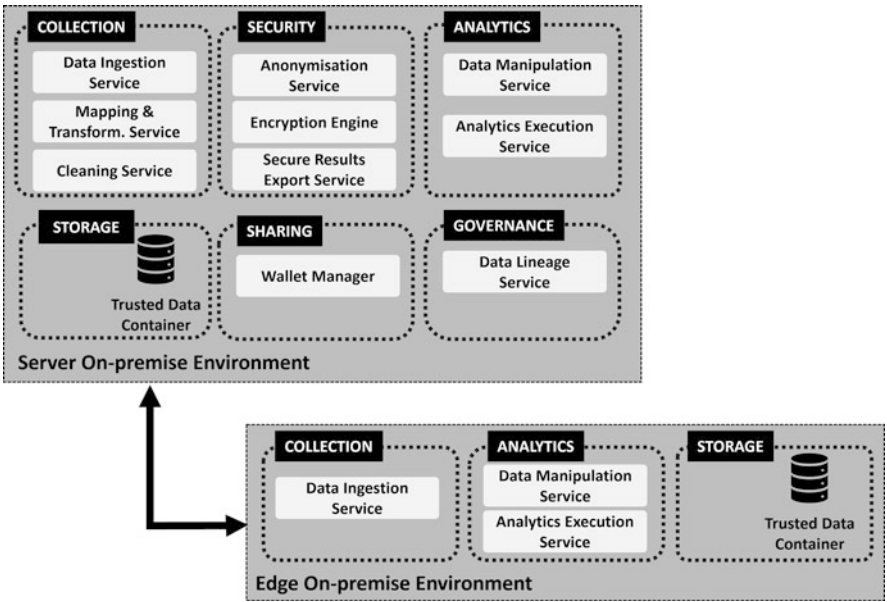

Fig. 5 Detailed component view of the SYNERGY On-Premise Environments

prior to uploading them in the SYNERGY Core Cloud Platform; (b) preparing and storing the own data assets "locally" in case they are not allowed to even leave a stakeholder's premises; and (c) running analytics "locally" over data that are also stored "locally."

As depicted in Fig. 5, according to the instructions received by the Master Controller in the SYNERGY Core Cloud Platform, a data check-in job is executed in the Server On-Premise Environment as follows: the **Data Ingestion Service** is

responsible for collecting the necessary data, the **Mapping and Transformation Service** for processing the data (to ensure their alignment with the CIM), the **Cleaning Service** for increasing the data quality, the **Anonymization Service** for handling any personally identifying or commercially sensitive data, and the **Encryption Engine** for encrypting the data. Then, the data are either stored locally in the Trusted Data Container or transferred to the SYNERGY Core Cloud Platform where they are permanently stored. It needs to be noted that in case an active smart data asset contract's terms allow it, the data assets that have been acquired by an organization can be also downloaded in the Server On-Premise Environment to be used to complement an analysis, again through the Data Ingestion Service, and shall be decrypted with the help of the Encryption Engine.

In order to execute a data analysis job "locally" in the Server On-Premise Environment, the Master Controller of the SYNERGY Core Cloud Platform appropriately invokes the **Data Manipulation Service** and the **Analytics Execution Service** to run all necessary steps of the designed pipeline. The results are stored in the Trusted Data Container and can be securely extracted from the Secure Results Export Service of On-Premise Environment Server Edition.

The **Wallet Manager** allows the organizations that have installed the On-Premise Environment Server Edition to securely handle the ledger account and the cryptocurrency funds of their organization. It is practically used to send payments for smart asset data contracts that allow an organization to buy data, but also to receive reimbursement for data assets that have been sold by the organization (especially in the context of a multi-party smart asset contract). The Data Lineage Service again allows a better view of the data asset's provenance.

The *Edge On-Premise Environment* has limited functionalities in respect to the Server On-Premise Environment due to the limited compute, memory, and storage capacity it can leverage in any gateway. It has a light version of (a) the Data Ingestion Service to ensure that a gateway may collect data as part of a data check-in job that has been configured in the SYNERGY Core Cloud Platform and (b) the Data Manipulation Service and the Analysis Execution Service that may run limited data analysis pipelines with restrictions.

4   Discussion

During the design of the SYNERGY Reference Architecture and iterative discussions performed in different technical meetings, the need to bring different stakeholders on the same page with regard to certain core end-to-end functionalities of the SYNERGY platform emerged. To this end, the basic workflows that the SYNERGY Cloud Infrastructure and On-Premise Environments will support from the user-oriented perspective of data asset providers and data asset consumers were designed and extensively discussed in dedicated workshops, focusing on the main challenges that are expected to be encountered:

- The **data check-in workflow (I)** allowing data asset providers to make available their data in the SYNERGY Energy Big Data Platform and the AI Analytics Marketplace.

    - *Challenge I.1*: Complexity of a fully flexible data check-in job configuration vs user friendliness. There is an explicit need for guidance and for setting certain quality thresholds in order to properly configure all steps since the settings for data mapping, cleaning, and anonymization cannot be fully and automatically extracted, but instead always have to rely on the expertise of the data provider who is uploading the data.
    - *Challenge I.2*: Performance vs security trade-off. When executing demanding pre-processing steps like Mapping, Cleaning, Anonymization, and especially Encryption over a dataset, certain restrictions need to apply (to avoid ending up with inconsistent data in a datastore), while real-time access to the processed data cannot be guaranteed. Increased security requires data replication and decryption in the different secure spaces of the data consumers, which cannot be instantly completed either.
    - *Challenge I.3*: Data profiling completeness vs status quo. In order to facilitate search, full profiles of different datasets need to be provided which requires significant attention by a data provider. Data licenses profiling in particular appears as a pain-point in an industry who is not used in sharing their own data. Although fine-grained access data access policies are considered as instrumental in ensuring the business interests of the demo partners toward their competitors, their configuration needs to be straightforward explaining the exact implications.

- The **data search and sharing workflow (II)** allowing data asset consumers to find data of interest in the SYNERGY Energy Big Data Platform and the AI Analytics Marketplace and acquire them in a trustful and reliable manner based on smart data asset contracts.

    - *Challenge II.1*: Search performance over Big Data vs the metadata of encrypted data. Search functionalities need to be always adapted to different cases of how and where the data are stored and indexed.
    - *Challenge II.2*: Multi-party contracts as a necessity vs a necessary "evil." In order to properly handle the chain of licenses and IPR that are associated with analytics results that can be traded in the marketplace, the SYNERGY platform needs to act as a "man in the middle" that creates bilateral contracts with the data asset consumer and each involved data asset providers under a broad multi-party contract. Facilitating the data asset consumer in this case comes at the cost of complexity on the platform side. In order to properly handle multi-party contracts, payments over a cryptocurrency (supported by SYNERGY) are also enforced which may lower the entry barrier for the potential stakeholders, but also potentially decrease their trust.
    - *Challenge III.3*: Limitations on data access and retrieval. Retrieval of appropriate data assets is not contingent only on the existence of an active data

asset contract, but also on the actual location of the data (cloud vs on-premise environment of the data provider) and the terms that dictate the data transfer. Although cloud presence of unencrypted data ensures that they can be retrieved via user-defined retrieval queries, encrypted data and on-premise data can be potentially (if there is a provision for offline/off-platform storage) only retrieved as full files through the SYNERGY platform APIs.

- The **data analytics workflow (III)** allowing data asset providers and consumers to run analytics over their own and the acquired data assets in the SYNERGY Energy Big Data Platform and the AI Analytics Marketplace and gain previously unattainable insights.

  – *Challenge III.1: Pipeline configuration for a business user vs a data scientist.* When trying to design a solution that allows the design of analytics pipelines, different perspectives need to be considered: the perspective of a business user who needs to easily create pipelines and gain insights over data and the perspective of data scientists that expect more advanced functionalities for feature engineering, model training, and evaluation.
  – *Challenge III.2*: Customizable pipelines for basic vs pre-trained energy data analytics algorithms across different execution frameworks (ranging from Spark and Python/sk-learn to TensorFlow over Keras). Since the input data to run an analytics pipeline are available as uploaded by their stakeholders, they need to be easily manipulated through an interactive user experience in order to be fit as input to an ML/DL model.
  – *Challenge III.3*: Data and model versioning affect the execution of any analytics pipeline. The expected impact on performance in "real-time" data and analytics when the data are originally stored in an encrypted form or only on premise (with limited resources) cannot be disregarded.
  – *Challenge III.4*: Running analytics with data that are never allowed to leave their provider's premises (according to the applicable data asset contract terms) render secure multi-party computations as a necessity (despite their inherent limitations in terms of analysis richness).

It needs to be noted that the aforementioned challenges represent an extract of the challenges identified during interactive workshops in which the technical partners were requested to discuss the technical challenges they expect to be associated with each feature/requirement and comment on their technical feasibility, according to their experience. In parallel, the different end users (across five demo countries) were requested to evaluate (a) the actual importance/added value for own organization (by rating the importance and added value of the specific feature for their business operations) and (b) the perceived importance/added value for the electricity data value chain (by rating the importance and added value that they perceive the specific feature brings to their stakeholder category and the overall electricity data value chain). For the assessment, a scale between 1 (little or no added value/importance/impact) and 5 (extremely high added value/importance/impact) was put into use, as indicatively depicted in Fig. 6.

Step	Functionality / Feature	Actual Importance / Added Value for own Organization	Perceived Importance / Added Value for the electricity data value chain
I-1. Data Ingestion Configuration	Upload of file(s) (csv, json, xml, other) that have been already extracted by the energy stakeholder's back-end systems	4.31	4.15
	Data ingestion via energy stakeholders' APIs	4.00	4.23
	Data ingestion via Open Data APIs (e.g. ENTSO E Transaparency Platform)	3.62	3.92
	Data ingestion via the SYNERGY Platform APIs	4.38	4.00
	Streaming Data Ingestion via the SYNERGY Core Cloud Platform (SYNERGY PubSub mechanism)	3.25	3.33
	Streaming Data Ingestion via stakeholders' own (PubSub) mechanisms	2.75	3.58
I-2. Data Mapping & Harmonization Configuration	Automatic mapping predictions from the data sample to the SYNERGY Common Information Model	4.31	4.31
	Easy navigation and search for missing concept mappings in the SYNERGY Common Information Model		
	Manual updates and corrections of mappings from the data sample to the SYNERGY Common Information Model		
	Manual definition of transformation rules (for measurement units, datetime format and timezones, etc)		
	Manual definition of advanced transformation rules (e.g. for merging data columns to comply with the SYNERGY Common Information Model)		
	Semantic enrichment of the data concepts with machine-processable information (e.g. information from codelists, vocabularies)		
	Manual suggestions for missing concepts in the SYNERGY Common Information Model		
I-3. Data Cleaning Configuration	(Semi-)Automatic quality check of the data and assessment of quality level	4.31	4.31
	Recommendation for cleaning rules to improve data quality		
	Definition of cleaning rules (for handling outliers, for handling missing data values, etc.)		
I-4. Data Anonymization Configuration	Semi-automatic check for identifying the "privacy-risky" columns of a dataset	4.31	4.31
	Definition of anonymization rules for "identifying", "quasi-identifying" and "sensitive" columns/variables		
I-5. Data Encryption Configuration	Encrypt the whole dataset	4.08	4.23
	Encrypt manually-selected "columns" of the dataset	3.67	3.92
	No encryption is needed for the dataset	2.31	2.15
I-6. Data Storage Configuration	Upload in the SYNERGY Core Platform	3.92	4.00
	Instantly replicate in my organization's Secure Experimentation Playground in the SYNERGY Core Platform	3.62	3.77
	Store in my organization's On-Premise Environment	4.23	4.23
I-7. Metadata Enrichment	Detailed data profiling of the data to be uploaded according to the SYNERGY metadata schema	3.33	3.33
	Quick check-in for confidential data assets (without filling in all metadata)		
I-8. Data License Definition	Detailed license profiling of the data to be uploaded (e.g. IPR, terms & conditions, pricing)	3.77	3.92
	Step-by-step guidance on how to define the appropriate license of the data to be uploaded	3.62	3.62
	Creation of a "license" contract between the data provider and the SYNERGY platform	4.08	4.08
I-9. Data Access Policies Definition	Definition of access policies regarding who is permitted to access the data (e.g. Organizations from Greece can access the data)	4.23	4.38
	Definition of access policies regarding who is denied access to the data (e.g. No DSO can access the data)	4.31	4.46
I-10. Data Collection Execution	Data uploading in SYNERGY Core Platform through the On-Premise Environment, according to the configuration provided	3.83	3.83
	Data uploading in SYNERGY Core Platform directly, according to the configuration provided		
	Data uploading in the On-Premise Environment only, according to the configuration provided		
	Automatic and on-demand data transfer to the organization's Secure Experimentation Playground		
	Scheduled execution of API data retrieval (e.g. every minute/ hour/ day, at defined schedules)		
	Real-time execution of streaming data retrieval		

Fig. 6 SYNERGY platform features assessment by the demo partners for own and industry business value for " data check-in" for a data provider

5 Conclusions

This paper focused on the motivation and state of play behind energy data platforms, marketplaces, and essentially Data Spaces in order to present the SYNERGY Reference Architecture that aims to facilitate electricity data value chain stakeholders to (a) attach value to their own data assets; (b) gain new insights over their data assets; (c) share and trade their own data assets in a trustful, legitimate manner; and (d) enjoy the benefits of the reuse of their own data assets. The different layers of the architecture as well as the different components across the SYNERGY Data Services Bundles have been elaborated, while the core technical challenges have been introduced.

The next steps of our work include the finalization of the beta release of the SYNERGY integrated platform (which is currently on its alpha, mockup version) and its demonstration and use by different electricity data value chain stakeholders.

References

1. CapGemini. (2015). *Big Data BlackOut: Are utilities powering up their data analytics?*.
2. Bhattarai, B. P., et al. (2019). Big data analytics in smart grids: State-of-the-art, challenges, opportunities and future directions. *IET Research Journals*, 1–15.
3. Daneshmand, M., et al. (2017). Big challenges for big data in the smart grid era. *ECN Magazine*.
4. Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (2017). *European big data value strategic research & innovation agenda*. Big Data Value Association.
5. Zillner, S., Bisset, D., Milano, M., Curry, E., García Robles, A., Hahn, T., Irgens, M., Lafrenz, R., Liepert, B., O'Sullivan, B., & Smeulders, A., (eds) (2020). *Strategic research, innovation and deployment agenda - AI, data and robotics partnership. Third Release*. September 2020, Brussels. BDVA, eu-Robotics, ELLIS, EurAI and CLAIRE.
6. HBR. (2006). *Strategies for two-sided markets*. https://hbr.org/2006/10/strategies-for-two-sided-markets
7. Biliri, E., Pertselakis, M., Phinikettos, M., Zacharias, M., Lampathaki, F., & Alexandrou, D. (2019). Designing a trusted data brokerage framework in the aviation domain. *PRO-VE, 2019*, 234–241.
8. Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: An initial survey. *ACM SIGMOD Record, 42*(1), 15–26.
9. Stahl, F., Schomm, F., Vomfell, L., & Vossen, G. (2015). *Marketplaces for digital data: Quo vadis? (No. 24)*. Working Papers, ERCIS-European Research Center for Information Systems.
10. Stahl, F., Schomm, F., Vossen, G., & Vomfell, L. (2016). A classification framework for data marketplaces. *Vietnam Journal of Computer Science, 3*(3), 137–143.
11. Koutroumpis, P., Leiponen, A., & Thomas, L. D. (2017). *The (unfulfilled) potential of data marketplaces (No. 53)*. ETLA Working Papers.

12. EURELECTRIC. (2016). *The power sector goes digital - Next generation data management for energy Consumers.* https://www.eurelectric.org/media/2029/joint_retail_dso_data_report_final_11may_as-2016-030-0258-01-e.pdf

13. NordREG. (2018). *Implementation of data hubs in the Nordic countries.* http://www.nordicenergyregulators.org/wp-content/uploads/2018/08/NordREG-Status-report-on-datahubs-June-2018.pdf

14. European Distribution System Operators. (2018) *E.DSO Policy Brief on Open Data.* https://www.edsoforsmartgrids.eu/wp-content/uploads/EDSO-Open-Data-Policy-Brief_1812_final-1.pdf

15. BDVA. (2017). *European big data value - Strategic research and innovation agenda, Version 4.0.* Retrieved from https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf. Accessed December 30, 2020.

16. CEN-CENELEC-ETSI Smart Grid Coordination Group. (2012). *Smart grid reference architecture.* Retrieved from https://ec.europa.eu/energy/sites/ener/files/documents/xpert_group1_reference_architecture.pdf. Accessed December 30, 2020.

17. EC. (2011). *Smart Grid Mandate Standardization Mandate to European Standardisation Organisations (ESOs) to support European Smart Grid deployment (M/490 EN).* Retrieved from https://ec.europa.eu/energy/sites/ener/files/documents/2011_03_01_mandate_m490_en.pdf. Accessed December 30, 2020.

18. Bröring, A. (2019). *Future data marketplaces.* https://www.bdva.eu/node/1220.