Virtual Energy System

Common framework

Demonstrator technology review
March 2023

Contents

Executive summary

Recommendations from the demonstrator technology review

Background

ESO have launched the VirtualES programme to enable the creation of an ecosystem of connected digital twins of the entire energy system of Great Britain, which will operate in synchronisation to the physical system. It will include representations of electricity and gas assets and link up to other sectors.

Through research, expert interviews, and industry-wide engagement, 14 key socio-technical factors were identified which are considered necessary for the development and delivery of the VirtualES today.

Following the example set by the National Digital Twin Programme and the Digital Twin Hub through their Climate Resilience Demonstrator project (CReDo), the VirtualES is developing a demonstrator that is focused on a *whole-system flexibility* use case.

This document contributes to the development of this demonstrator, currently being progressed through an NIA-funded project in Alpha phase. Its purpose is to establish that it is possible to make energy data visible, accessible, and shareable to actors across the industry through a secure and scalable solution; and determine the High Level Design for that solution.

Approach

This report assesses, evaluates and proposes data sharing solutions for the demonstrator. This was derived through desk-based research, stakeholder interviews, and discussions with platform providers - both within the energy sector and cross-sector.

It considers the various technology options and their combination, and provides recommendations on the technology architecture for both the demonstrator and the future VirtualES vision.

This report should be read in conjunction with the demonstrator data needs & gaps report, which outlines the data types required to be shared using the technology described in this report.

Conclusions

We recommend that:

- A fully distributed architecture for the VirtualES is considered. This will likely involve a combination of technological solutions to accommodate multiple future use cases, and to meet a range of technology, data, and security & governance requirements.
- The design should adopt Data Mesh principles to deliver an architecture suited for decentralised data sharing.
- The technology stack for the demonstrator will need to both fulfil the requirements of the demonstrator, and also set the foundation for future developmental iterations of the VirtualES to build towards a fully distributed architecture.
- Open Energy is considered a suitable data portal for the demonstrator use case.

Nomenclature

Nomenclature

AAR - Automatic Asset Registration
ACID - Atomicity, Consistency, Isolation, Durability
API - Application Programming Interface
CDC - Change Data Capture
CIM - Common Information Model
CKAN - Comprehensive Knowledge Archive Network
CPNI - Centre for the Protection of National Infrastructure
CReDo - Climate Resilience Demonstrator
DAFNI - Data & Analytics Facility for National Infrastructure
DAP – Data Access Portal
DDoS - Distributed Denial of Service
DNS - Domain Name System
DT- Digital Twin
EDT - Energy Data Taskforce
EDiT - Energy Digitalisation Taskforce
ETL - Extract Transform & Load
HLD - High Level Design
IoT - Internet of Things
Ofgem - Office of Gas and Electricity Markets
NCSC - National Cyber Security Centre
NDTP - National Digital Twin Programme
NeRDA - Near Real Time Data Application
NIS - Network & Information Systems
SQL - Structured Query Language
SSO - Single Sign-On
TCP - Transmission Control Protocol
VirtualES - Virtual Energy System
VPN - Virtual Private Network

Terms

- **Data Producers** – this refers to organisations and users of the VirtualES (e.g. DNOs, TNOs etc.), that will share their data externally with other organisations through the VirtualES. There could also be provision of data from other sectors of use to energy sector actors via the VirtualES.
- **Data Consumers** – this refers to organisations and users of the VirtualES that will consume data from external organisations (i.e. producers) through the VirtualES. These organisations may not necessarily be in the energy sector but could be innovators and technical sectors e.g. transport.
- **Digital Spine** – according to EDiT, the Digital Spine is a network of connected nodes deployed by organisations across the energy sector enabling a thin layer of interoperability and interaction as each node ingests, standardises and shares energy system data across all players.

1

Approach

Context

What is the Virtual Energy System?

The Virtual Energy System

The ambition of the Virtual Energy System (VirtualES) programme is to enable the creation of an ecosystem of connected digital twins of the entire energy system of Great Britain, that will operate in synchronisation to the physical system. It will include representations of electricity and gas assets and link up to other sectors.

This ecosystem of connected digital twins will enable the secure and resilient sharing of energy data across organisational and sector boundaries, facilitating more complex scenario modelling to deliver optimal whole-system decision making. These whole-system decisions will result in better outcomes for society, the economy, and environment by balancing the needs of users, electricity and gas systems and other sectors.

Creating the VirtualES is a socio-technical challenge that requires a collaborative and principled approach, aligned with the National Digital Twin Programme, and other energy sector digitalisation programmes.

The VirtualES is delivered through three workstreams:

- Workstream 1 - Stakeholder engagement
- Workstream 2 - Common framework & principles
- Workstream 3 - Use cases

Workstream 2 - Common Framework & Principles

This report forms part of workstream 2.

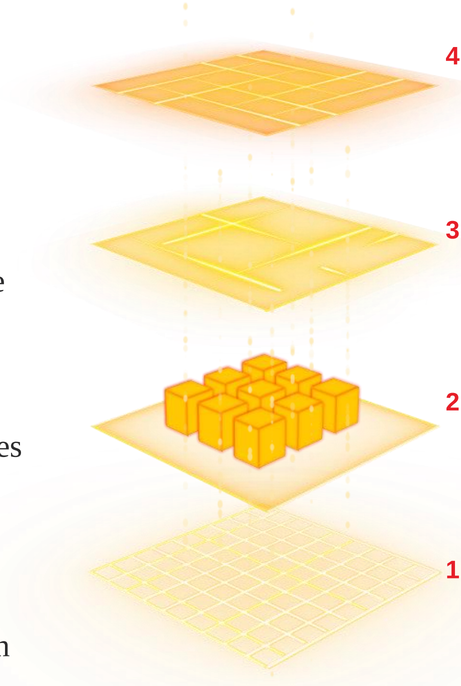
The objective of this workstream is to develop the socio-technical common framework that will form the foundation of the VirtualES – enabling the creation of this ecosystem of connected digital twins.

Through research, expert interviews, and industry-wide engagement, 14 key socio-technical factors were identified which are considered necessary for the development and delivery of the VirtualES today.

These 14 identified factors are grouped by the categories of People, Process, Data, and Technology**.** Six of these factors were prioritised based on their potential impact on the VirtualES objectives and their relative maturity across the wider energy sector.

Following the example set by the National Digital Twin programme and the Digital Twin Hub through their Climate Resilience Demonstrator project (CReDo), this workstream is now developing a demonstrator that is focused on a *whole-system flexibility* use case.

This document contributes to the development of this demonstrator, currently being progressed through an NIA-funded project in Alpha phase.



4. The data becomes more layered, these interactions will create valuable insight to help guide and govern how we generate, manage, store, and consume energy.
3. Populated by existing and new digital twins – replicas of physical components of our energy system
2. Each digital twin will contribute to and access real-time data on the status and operation of other elements of the system
1. A social-technical common framework, with agreed access, operations and security protocols

Virtual Energy System

Indicative components of the Virtual Energy System

Developing a common framework

Published research and reports for the common framework

Throughout the development of the common framework, the approach has been industry-led, consultative, and collaborative.

This approach, coupled with explicit and proactive engagement within the energy sector and with cross-sector stakeholders, is necessary for the successful development of the common framework, delivery of the VirtualES, and ultimately in achieving sector-wide adoption.

All work has been conducted openly, with the six reports completed to date all published online.

Following the SIF Discovery project (report #3), the demonstrator was further developed using the whole-system flexibility use case (report #4).

The demonstrator is currently progressing through an NIA-funded project in Alpha phase, and is being delivered in line with the project plan (report #6).

1. External benchmarking

Understanding the cross-sector and global best practice for connecting assets, systems, and digital twins.

Read the report

2. Defining the common framework

Determining the key socio-technical factors that need to be considered for the VirtualES to succeed.

Read the report

3. Demonstrating the common framework

Collaboratively prove and demonstrate, with industry, how the socio-technical principles work.

This was a Round 1 SIF Discovery project.

Read the report

4. Whole system flexibility use case definition

Further define the “whole-system flexibility” use case that is recommended as the initial use case to demonstrate the common framework.

Read the report

5. Demonstrator data standards, data portals, and data licensing

Identified data standards and outline data licensing considerations applicable to the use case. Initial review of currently available public energy sector ‘data portals’.

Read the report

6. Demonstrator project plan & advisory groups

Proposed delivery plan, governance structure, advisory groups approach, and cross-workstream collaboration that will enable the successful delivery of the demonstrator.

Read the report

Delivery team

Supporting the development of the social-technical common framework

The development of the common framework has been delivered by Arup and supported by the Energy Systems Catapult and Icebreaker One. It has been sponsored by the Electricity System Operator (ESO) and National Gas Transmission (NGT) through the Network Innovation Allowance (NIA).

The purpose of the RIIO-2 NIA is to provide funding to Gas Transporter and Electricity Transmission Licensees to allow them to carry out innovative projects, that focus on the energy system transition or addressing consumer vulnerability, which are outside of business-as-usual activities.

- **Electricity System Operator (ESO):** ESO is responsible to ensure a reliable, secure system operation to deliver electricity when customers need it. ESO balances the supply and demand on the system day to day, second by second, and coordinates with networks to transfer electricity from where it is generated to where it is needed.
- **National Gas (NGT):** National Gas own and operate the national gas network in addition to maintaining and managing the 7,000,000 domestic industrial and commercial combined gas assets around the UK.
- **Arup:** An employee owned, multinational organisation with more than 15,000 specialists, working across 90+ disciplines, with projects in over 140 countries and the mission to ‘shape a better world’. Arup have extensive energy and cross-sector digital twin expertise, actively contributed to the National Digital Twin programme, and are members of the Digital Twin Hub.
- **Energy Systems Catapult (ESC):** An independent, not-for-profit centre of excellence that bridges the gap between industry, government, academia, and research. Set up to accelerate the transformation of the UK’s energy system and ensure businesses and consumers capture the opportunities of clean growth. ESC are responsible for the Energy Data Task Force (EDTF) & Energy Digitalisation Task Force (EDiT).
- **Icebreaker One (IB1):** An independent, non-partisan, non-profit organisation with a mission to ‘make data work harder to deliver Net Zero’ by creating open standards for data sharing across agriculture, energy, transport, water, and the built world.

Together the five organisations assembled a delivery team to effectively collaborate and deliver the objectives of this workstream.

Introduction

Purpose of this document

Purpose

This document presents the findings of **WP2.2 - Technology**, developed as part of the common framework demonstrator Alpha phase.

Its purpose is to establish that it is possible to make energy data visible, accessible, and shareable to actors across the industry through a secure and scalable solution. This document contains the following deliverables:

- Technology review report (M3)

Summary

The success of the VirtualES depends on the implementation of a suitable data sharing architecture that allows for various stakeholders within the energy sector to securely and effectively share their data.

Stakeholders must have the ability to control access to their data, as well as the ability to determine what data they can share, whilst ensuring that the system does not pose a risk to their own systems or infrastructure from cyber attacks.

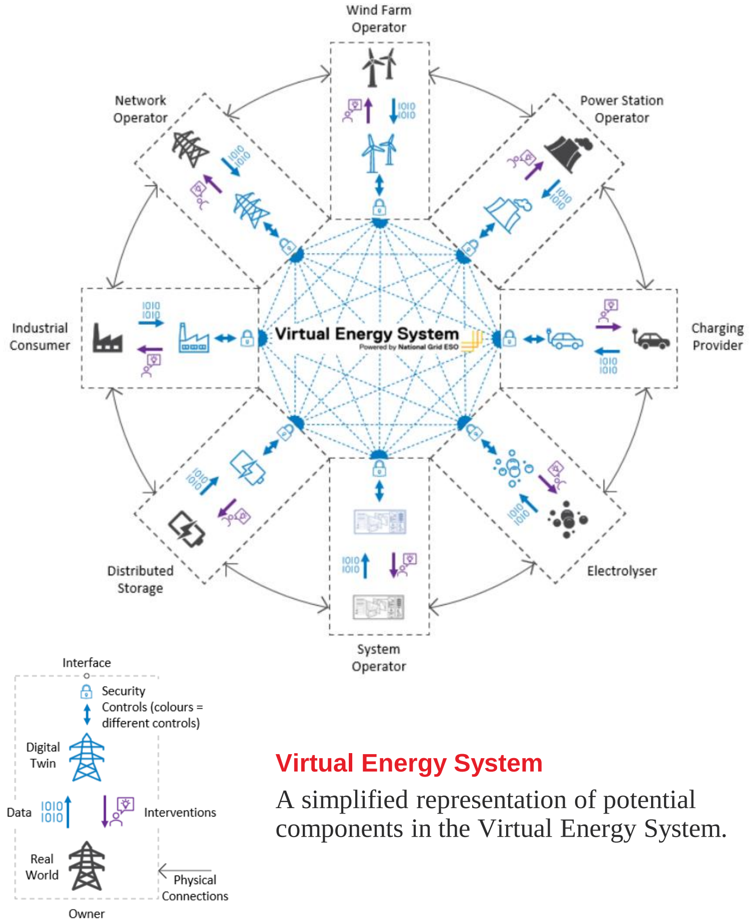
This document outlines key data, security and technology considerations for data sharing. These considerations were used to assess various data sharing options and their technical implementation.

A review of a selection of existing data sharing platforms was conducted to understand approaches from different sectors, and assess their suitability for adoption.

These findings were then translated into a High Level Design (HLD) for the demonstrator use case and also the future VirtualES vision.

Our approach shares similarities with the Digital Twin Hub Climate Resilience Demonstrator (CReDo), whereby a single use case was used to demonstrate the advantages of combining data and insights across sectoral and organisational boundaries, through the creation of an ecosystem of connected digital twins.

The VirtualES demonstrator aims to showcase and test many elements of the VirtualES socio-technical common framework, including the priority factors such as developing an interoperable technology stack and increasing data visibility and enabling sharing.



Virtual Energy System

A simplified representation of potential components in the Virtual Energy System.

Methodology

Identifying and evaluating the data sharing options for the VirtualES demonstrator

Approach overview

The methodology consists of a five-step process that assessed, evaluated and proposed data sharing solutions for the demonstrator.

1. **Review of existing literature and data:** Establish the baseline information and data available for the proposed use case, including data, technology and security considerations.

   This review of literature and data was informed by recent reports from the Energy Data Task Force (EDTF), Energy Digitalisation Task Force (EDiT) and Ofgem; security guidelines from NCSC, UK CPNI, DNI, and the published research and reports for the VirtualES common framework.

2. **Desk-based research**: Conduct desk-based research, to gather information around data sharing options and technical implementation considerations.

   This research was conducted using literature, blogs, and articles pertaining to architectural best practices for storage and distribution of data.

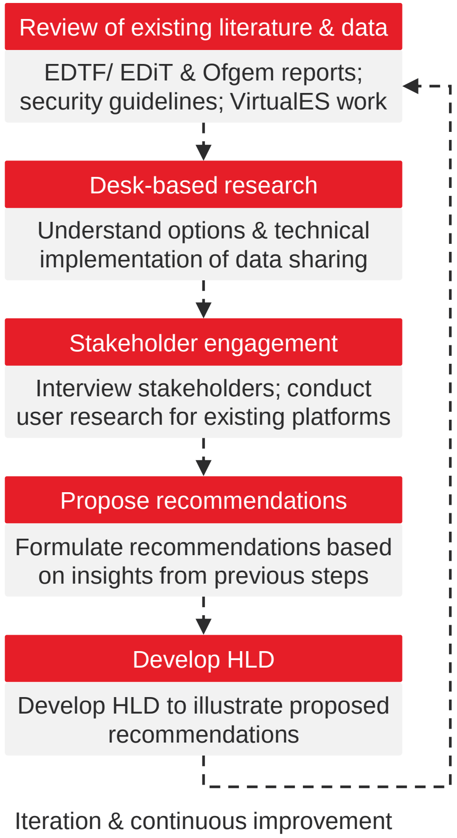
3. **Stakeholder engagement:** Conduct user research and interviews with relevant stakeholders and platform owners to establish an understanding of the current data sharing landscape.

   Interviews were conducted with platform owners for GreenSync, Open Energy, and Telicent, in addition to interviews with 16 key stakeholders, representing data governance leads, architects and planning roles across both electricity and gas networks. These interviews provided insights into data sharing options and user requirements. The stakeholders developed our understanding of current data sharing options and potential application of existing platforms.

4. **Propose recommendations**: Based on the insights and conclusions gained from the previous three steps, formulate recommendations for potential data sharing options for the demonstrator.

5. **Develop High Level Design (HLD):** Develop a HLD to illustrate the proposed recommendations.

   The recommendations and HLD were formulated based on the insights, assessments and conclusions from the research and stakeholder engagement.



2

Data, technology, and security recommendations and considerations

Data sharing considerations

Data sharing recommendations and considerations applicable to the VirtualES

Overview

The data which is required to be shared within the VirtualES demonstrator is outlined in the data needs assessment report.

This section considers how data should be shared within the sector, taking guidance from recommendations made in recent reports from the EDTF, EDiT and Ofgem.

It outlines each aspect of data management which needs to be considered, and the rationale for the positives and negatives for each data sharing option.

The summarised recommendations include data governance requirements that are fundamental for the VirtualES.

Recommendations from energy sector guidance

Energy Data Task Force / Digitalisation Task Force

- Data should be discoverable, searchable, and understandable. (EDTF).
- Data should have common structures, interfaces and standards. (EDTF).
- Data should be secure and resilient. (EDTF).
- Adopt digital security measures fit for purpose for a zero trust principle and a sharing culture (EDiT).

Ofgem's Data Best Practice guidance

- Potential data users can identify the data assets of data custodians, and pursue access to these data sets.
- Data custodians must ensure that the metadata associated to data assets is discoverable to data users, subject to the outcome of an open data triage process.
- Data owners must ensure data assets are interoperable with data assets from other data and digital services.
- Licensees must make data available in such a way that it is reasonably easy for data users to:
  - Exchange data assets between systems.
  - Interface with data assets in licensee's systems.
  - Join data assets with other data assets, such as by using standard interfaces, standard data structures and/or common reference data.

The following considerations should be made

Data governance & management

- Data stewardship and ownership.
- Data classification and retention policies.
- Data trustworthiness.
- Data discovery, cataloguing and publishing.
- Clear data ethics policies.

Data standards/models & sharing

- Open data standards/models e.g. CIM.
- Interoperability between various sector actors.
- Common principles for data management.
- Appropriate governance working groups/committees.
- Availability and documentation of data models & schemas for consumptions and extensions.
- Sharing of best practices across the wider sector.

User experience

- Consistent data standard, presentation & accessibility.
- Well understood and accessible metadata so consumers can understand content of data.
- Consistent formats (e.g. CSV, JSON, XML etc.).
- Ease of data accessibility, search, discovery, access.

Security considerations

Security recommendations and considerations applicable to the VirtualES

Overview

Data security is a crucial concern for the VirtualES, where sensitive information, such as power grid operations and energy consumption data, must be protected from potential cyber threats. To ensure protection and integrity of the data, the VirtualES should be aligned to appropriate security standards and practices.

It is important to note that this is not solely a technical challenge, but the combination of data, people, and processes also need consideration - ensuring appropriate controls and governance processes are in place.

Recommendations from security standards

- Continuous alignment to international security standards and best practices.
- Sector specific protocols that emerge over time.
- Cyber security in accordance with NCSC guidelines, NIS directive and ISO 27001.
- Data privacy principles standards for IoT implementation in accordance with the UK CPNI.
- Enforcing clear trust protocols (such as around authorisation) for data usage and integrity.
- Alignment with the developing principles of the National Digital Twin Programme.

Security considerations

Security architecture

- Microservice architecture and system isolation (isolation of web, application and database layers).
- Monitoring, observability and auditing e.g. API calls, user and application activity.
- Security automation: controls managed as code.
- Response routes for security incidents: accident management process; root cause analysis; threat detection.
- Web security: DDoS attacks; SQL injection attacks.
- Application and operating system hardening.
- Network security: security rules; virtual private clouds; routing configuration; DNS.
- Layered security: network layer security; firewalls; antivirus software; security for serverless applications.
- Secure API calls using TCP (for example).
- Cryptography protocols and immutability (long term view).
- Zero trust principle in accordance with NCSC.
- Global governance & security controls without using organisation or domain specific business logic.

Data security

- Data classification & categorisation.
- Data encryption (rest and transit).
- Legal & compliance considerations (including discovery, retention and deletion of data).
- Data observability & audit.
- Data licensing agreements.
- Data access control and logging.

Access control

- Authentication & Authorisation of users.
- Secure connections (SSO, VPN etc.).
- Federated identity management.
- User trust frameworks e.g. secure certificate exchange between trusted organisations.

Supply chain

The security and recommendations should also apply to the supply chain, and making sure appropriate data licensing agreements, standards, policies and access control mechanisms etc., are in place for organisations across the supply chain.

Technology considerations & requirements

Technology requirements and considerations applicable to the VirtualES

Overview

Technology selection for the VirtualES is a critical decision that requires careful consideration of various factors such as compatibility with existing systems, existing governance structures, complexity of implementation, and ease of scalability.

The chosen technology should also be able to scale to meet future demands and provide appropriate access control and data governance capabilities. Additionally, the quality of the user journey must be taken into account to ensure that the chosen technology provides a smooth transition during the change process.

Overall, the goal is to select a data sharing technology that can support the efficient and effective operation of the VirtualES. This section outlines the key considerations that will be used to evaluate the various data sharing options in Section 3.

The following considerations are for the long-term vision of the VirtualES. By considering the overall, high-level requirements for a fully distributed platform, it is possible to work backwards and pick a suitable solution for the demonstrator use case, with the aim of transitioning to a complete design through further development iterations.

Technology considerations

- Open-source software and open standards.
- Cloud hosted.
- Distributed storage and parallel processing.
- Event streaming to accommodate real time, high-volume data.
- Asynchronous data sharing i.e. producers share data even when consumers are offline, allowing consumption of data independently from producers.
- High fault tolerance: for the VirtualES platform and producers' data sharing pipelines. May require provisioning of replica services / compute instances.
- High availability.
- Ability to periodically poll for new data.
- Scalable big data storage (if storage is required for example for historical data).
- Ability to accommodate additional data systems and users coming online and ingestion of real-time data without any changes in the pipeline.
- Providers and consumers of data feeds can be any kind of data system (database, search system, cache, Hadoop etc.).
- Event-driven architecture.

User journey

- Data consumers should be able to easily find useful datasets, and data producers should be able to easily share data.
- Licensing and moving the data closer to consumers' existing data and analytics should be quick and easy by using pre-agreed licence terms, and openly-published access conditions.
- Consumers need a trusted application to securely and effectively find, access, subscribe to, and consume data from multiple sources.
- Producers should not have to build and maintain non-differentiating technologies to package, deliver and provide access to the data when delivering data to multiple consumers.
- Consumers should be notified when producers provide new data, thereby allowing consumers to automatically consume and use the data.
- Consumers should have the ability to provide feedback about a dataset when they are aware of data quality issues which will assist the data owner to improve the dataset.

Design considerations using a Data Mesh approach

Technology architecture considerations using Data Mesh principles

For any technology solution, considerations for the overall architecture and the use of modern design principles needs to be built in as part of the overall implementation.

A notable, and increasingly adopted design philosophy is Data Mesh, a paradigm shift from centralised architectures (typically data lakes or data warehouses), to one that draws from modern distributed architectures.

Data Mesh embraces ubiquitous data through the convergence of *Distributed Domain Driven Architecture, Self-serve Platform Design* and *Product Thinking with Data.*

- **Domain oriented data:** this is about shifting how data is shared. So instead of an organisation (or enterprise domain) flowing their data into a centralised platform for ingestion, processing and serving, the organisation will instead process, store and serve that dataset for access by others.

  This shifts the thinking from a push and ingest (traditionally through ETL), to a serve and pull model across organisations. This is a reverse for how to traditionally think about data, its locality and ownership.

- **Product thinking with data:** for a distributed data platform to work, organisations must apply product thinking with their data. This means that the data they offer up to consumers must have certain characteristics to make it fit for consumption e.g. trustworthy, self-describing, secure, inter-operable.

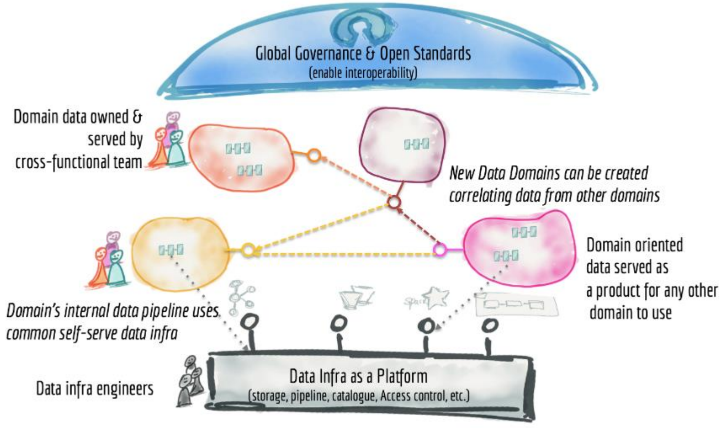
- **Self-serve platform design:** building a common data infrastructure that provides the tooling and techniques for organisations to set up their data pipelines, storage and streaming infrastructure. This data infrastructure will not contain any organisation/domain specific concepts or business logic i.e. keeping it domain agnostic; and it must abstract all the technology complexity away from the users, thereby offering the tools & capabilities in a self-serve manner.

  These capabilities may include: scalable big data storage, data encryption, data discovery and catalogue, data access control and logging, data quality metrics etc. Some of these elements may be met by the VirtualES.

To summarise, a Data Mesh design is a distributed data architecture that uses global governance and standardisation to drive interoperability, which is enabled by a self-serve data infrastructure.

This provides governing principles which supersedes traditional approaches. These principles include:

- Serving over ingesting.
- Discovering and using over extracting and loading.
- Publishing events as streams over flowing data around via centralised pipelines.
- Ecosystem of data products over centralised data platforms.



Source: Zhamak Dehghani, martinfowler.com

3

Data sharing options

Summary of options

Data sharing options considered for the demonstrator and future VirtualES

Overview

This section assesses and forms conclusions for six data sharing options. Each option contains a summary, positives, negatives, and conclusions for each technology architecture approach.

The assessment for each option uses the criteria of data, security, governance and technology considerations, and reviews suitability from the perspectives of both the demonstrator use case and the future VirtualES.

The following options are assessed:

- **Option 1:** point-to-point sharing
- **Option 2:** centralised datastore
- **Option 3:** data portal
- **Option 4:** distributed storage system
- **Option 5:** distributed streaming platform
- **Option 6:** data virtualisation

The assessment of the above options builds on the work conducted by the data sharing architecture industry collaboration group, particularly for options 1 to 4 where further detail on technical implementation and considerations are provided.

Summary of conclusion

A fully distributed architecture, that borrows principles from a data mesh design, coupled with a governance solution to provide assurance of and for organisations exchanging data, is envisioned for the VirtualES.

However, a distributed architecture contains many complex parts including distributed storage, processing, and communication. Furthermore, any single data sharing option may not meet all the requirements for all potential future use cases for the VirtualES.

Therefore, the future technology architecture may entail an amalgamation of technologies and data sharing options to satisfy multiple users with multiple needs.

The main conclusions for each data sharing option, and the options role in the future VirtualES is summarised below. Full details are given on the subsequent pages

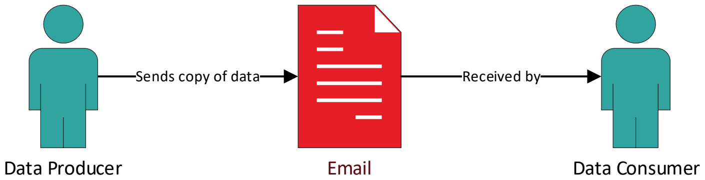
- **Option 1 - point-to-point sharing:** Direct sharing (via emails) is not recommended because it does not meet the range of technology, data, governance and security requirements.
- **Option 2 - centralised datastore:** Suitable for use cases where the data has to be centralised and administered by a central owner, and where, for example, historical data is required for analysis.
- **Option 3 - data portal:** Suitable when acting as the user interface and data catalogue for users to search and discover their data.
- **Option 4 - distributed storage system:** Suitable to distribute data amongst different nodes and servers for instances where performance, security, and large volumes of data are required, perhaps for analytical purposes.
- **Option 5 - distributed streaming platform:** Suitable for sharing of real-time data, and orchestration of data sharing with users and data stores by implementing an event-driven architecture.
- **Option 6 - data virtualisation:** Is not considered as a preferred option, due to the its complexity and uncertainty on the applicability of its architecture for the future VirtualES.

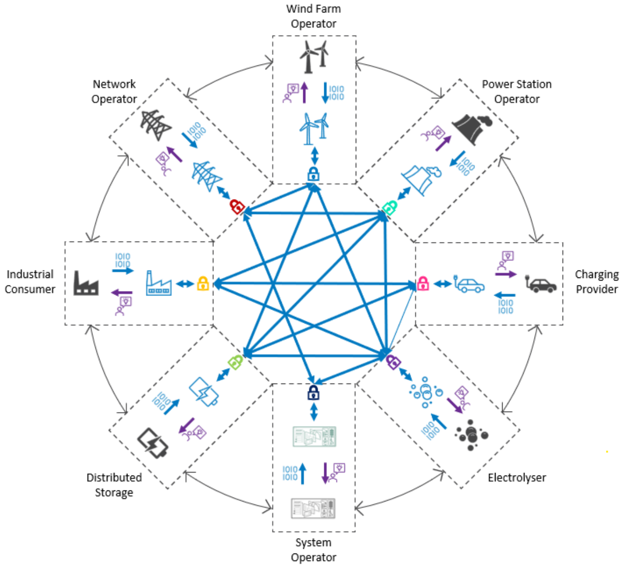
The HLD for the demonstrator use case should adopt a technology option(s) that can be used as the foundation for future iterations of development in order to realise a distributed architecture. This must be underpinned by a range of governance and security controls, to enable compliant and secure data sharing, and coupled with open standards to enable interoperability.

Option 1: point-to-point sharing [not recommended]

Description

A data producer will take a copy of the relevant data - which may be a CSV file or JSON/XML (CIM) and send the copy of the data to a data consumer. For example, via email or conventional postal services.





Positives

- **Convenience:** relevant data can be emailed or transferred whenever it best suits the producer & consumer.
- **Simplicity:** straightforward and minimises any additional constraints and processing steps to receive the data (i.e., data does not need to be uploaded to a central portal or a database before consumption etc.).
- **Ownership:** Ownership of data and control is maintained.
- **Permission simplicity:** the data producer chooses whom to provide the data to with relative ease, without needing to store in a central pot with a series of conditions and access control configurations.

Negatives

- **Long-term vision:** does not offer a long-term enduring solution for the VirtualES and what it is trying to achieve in terms of a common framework for data sharing.
- **Limitations:** solution does not meet the range of data, technology, security & governance requirements.
- **Real-time:** sharing of real-time (or near real-time) would be difficult.
- **Data standards:** there is no way of validating the data standard/models used.
- **Scaling:** does not scale well because if there are multiple consumers and producers then the number of n nodes gives n^2 connections.

Conclusion

Whilst this option offers the most simple and straightforward solution, it would not meet the range of technology, data and security & governance requirements for the demonstrator and subsequently the long-term vision for the VirtualES, partly due to scaling issues when data is shared between multiple consumers and producers.

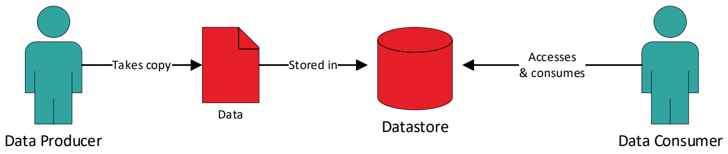
Option 2: centralised datastore [potential future application]

Description

A centralised and shared datastore is used to store relevant data and to connect the producers and consumers.

This will mean that the data producer will take a copy of the data, model it to store it in a database (relational, key-value, file format etc.) and makes it available for querying and consumption by a data consumer.

This could be used in instances where historical data is required for storage and consumption, and instances where the data does not necessarily have to be shared in real-time but, for example, through a batch process.



Positives

- **ACID compliance:** can offer ACID (atomicity, consistency, isolation, durability) compliance when storing and querying data.
- **Data consistency:** a centralised datastore can offer a single instance for the entire state making data consistency easier to achieve.
- **Performance:** can improve performance by creating secondary read-only replicas.
- **Near real-time sharing:** databases like Amazon DynamoDB and MongoDB can allow for users to share data in real-time. They can be accessed through secure APIs, which allows them to be integrated with other systems and applications.

Negatives

- **Performance:** if continuous processing with low delays is required, then querying/polling the database can become expensive and the amount of new data retrieved is lowered the more often the datastore is polled.
- **Ownership:** a central owner would be required to administer and manage the store.
- **Security:** security and governance controls could be difficult to implement if various organisations have different requirements when using the central store.
- **Data standards:** the data stored may be limited to data that conforms to a specific schema. If different models are required for sharing of data with multiple organisations then this could be restrictive.
- **Real-time data:** some traditional datastores do not cope well with real-time data. However, some of the cloud PaaS technologies could offer this capability.
- **Querying:** querying may become difficult and error-prone e.g., if the entire SQL code is required before deploying to a server.
- **Scaling (concurrent requests):** as new users come online, then the number of concurrent requests will also increase. This can make response times unpredictable if scaling is not possible or too expensive.
- **Scaling (concurrent updates):** if multiple users are performing updates concurrently then this can affect performance, unless a control mechanism is used to coordinate the concurrent updates.

Option 2: centralised datastore [potential future application]

Database technologies

Relational databases such as MySQL and PostgreSQL can be used for near real-time data sharing, but they typically require additional technologies or techniques to handle real-time updates. For example, using triggers to update multiple tables in response to a single change.

Some NoSQL databases, such as DynamoDB, are designed to handle high-performance, high-scalability, and near real-time data. These databases are often used in real-time applications such as gaming, social media, and IoT. They can also handle high-concurrency and large number of users.

In all the cases, it is important to ensure that the database can handle the expected load and concurrency of the application, and that it is properly secured and backed up to prevent data loss or corruption.

Conclusion

A centralised datastore could technically offer the minimum technology, data and security requirements for the demonstrator use case. Furthermore, some of the database technologies mentioned also provide good options for sharing of near real time data, with MongoDB being the database of choice for SSE's Near Real Time Data Application (NeRDA).

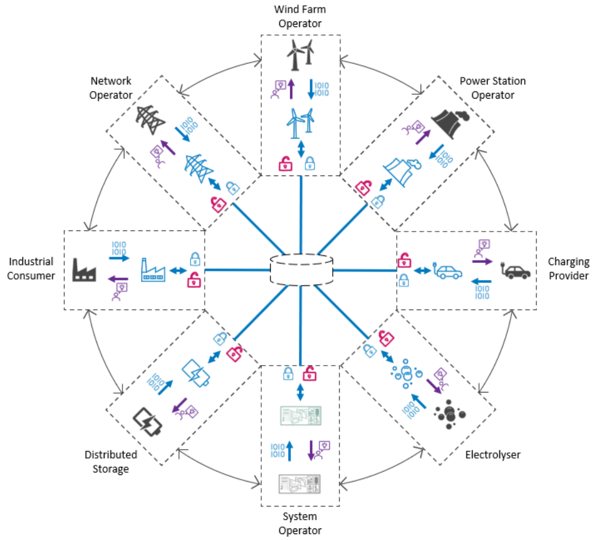
Considering the demonstrator use case will only involve sharing of data between a few organisations, some of the challenges around scaling would be mitigated. However, ownership and administration of the datastore would need to be decided, as centralised ownership would not provide a distributed architecture to match the long-term vision of the VirtualES.

Furthermore, a single centralised datastore would encounter challenges associated with scaling, polling, and handling of large volumes of real-time data.

Security and governance controls could also become a challenge if different organisations require different controls, and the sharing of common models could be limited if the datastore restricts the data model to one type of schema.

However, there may be use cases in the future where a centralised datastore would be useful. For example, the storage of historical data with aggregate views for simulation and analytical purposes between a select few producers and consumers, and for data warehouses.

Therefore, the future architecture of the VirtualES should have the ability to accommodate for such solutions as part of its overall design, but with a question around where the ownership of the datastore will sit.



Option 3: data portal [target for use case]

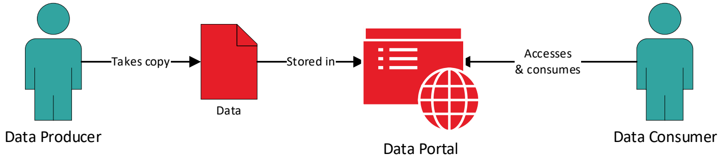
Description

A data portal is a web-based location where data and datasets are aggregated and published in any variety of machine and non-machine readable formats.

The intention is for data consumers and producers to search, discover, access and securely share energy data. A variety of these data portals currently exist, with different formats, metadata and datasets.

Some data portals host the data by using a database, whilst others (like Open Energy) does not host the data, but provides a cataloguing capability and directs users to where the data is hosted with the central administrator.

A high level assessment of existing data portals was conducted in previous VirtualES published reports. (Link)



Positives

- **Existing technology:** multiple data portals currently exist, with Open Energy having the most comprehensive list of metadata on different datasets. Furthermore, gaps and inefficiencies for data portals are well addressed.
- **Metadata management:** offers a data cataloguing solution for metadata management using established standards, thereby allowing users to search and find data.
- **Open data:** for open data, where access control is not a primary concern, then a data portal could offer this.
- **Security and governance:** data owners can choose who to share their data with and retain data in their own secure databases under their own management. Furthermore, they can impose the minimum compliance and governance controls to sufficiently address data quality and metadata management challenges.

Negatives

- **Consistency in quality:** existing data portals do not have consistent standards and requirements for publication, governance, presentation and availability of data.
- **Underlying technology:** the underlying technology for storage and consumption of data at a large scale may suffer some of the same limitations as that of a centralised datastore (Option 2). Some data portals may be suitable for some use cases and smaller datasets, however, this may not be appropriate to scale for a fully distributed architecture. Whilst it may provide a suitable cataloguing capability, the underlying storage, processing and distribution of real-time data will need further consideration.

Option 3: data portal [target for use case]

Conclusion

Existing data portals are suitable for data sharing where there are only a small number of participating organisations, relatively smaller volumes of data, and for open data which does not require real-time sharing.

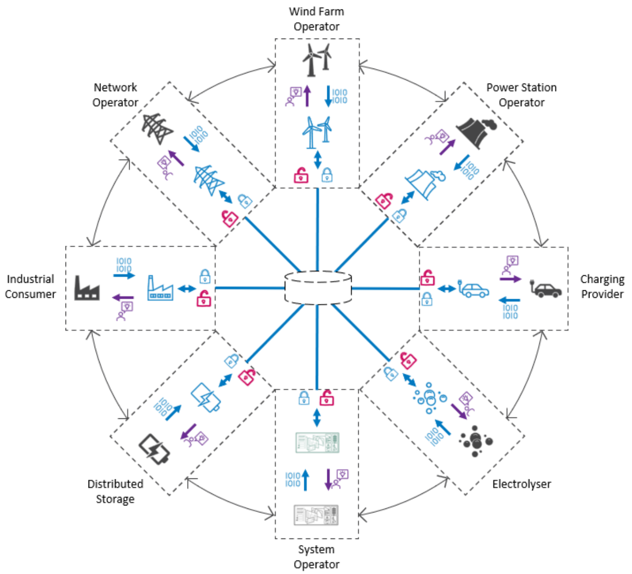
The Open Energy platform is an example which provides a comprehensive list of metadata on different energy sector datasets and offers a good data cataloguing capability. However, the platform does not orchestrate the sharing of data between producers and consumers.

Instead, the data is catalogued using CKAN and users are directed to where the data is stored under the management and administration of the data owners, often in their own organisational data portal.

Open Energy should be considered for adoption as part of the demonstrator use case, but it will need to be combined with a wider technology stack to enable distributed data sharing (section 4 provides a review of the platform and ways forward for adoption).

Data portals generally pose technology considerations for the underlying storage, processing and distribution of the data, as some existing data portals adopt centralised ownership which presents similar limitations as that of option 2.

In conclusion, a data portal offers a logical solution for the user interface and metadata management functionality for both the demonstrator and future VirtualES, however, it will need to be coupled with considerations for the storage, processing and sharing of data as part of a wider distributed architecture.

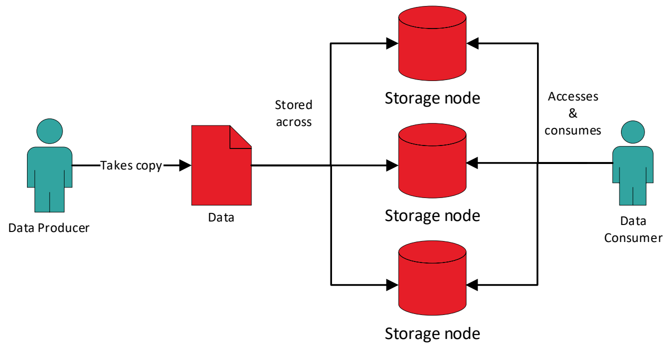


Option 4: distributed storage system [potential future application]

Description

A distributed storage system is infrastructure that can divide and split the data across multiple nodes/servers/data centres thereby allowing for massively scalable storage systems to be accessed by multiple consumers.

Popular examples of these are cloud storage technologies e.g. Amazon S3, Google Cloud Storage, Azure Blob etc. Another prominent example is Hadoop, an open-source framework which utilises clusters of servers to process and store data – typically used to manage big data.



Positives

- **Scalability:** distributed storage systems can be easily scaled horizontally (especially if using a cloud vendor) to accommodate larger volumes of data by adding more nodes to the system.
- **High availability:** with multiple nodes, the storage system can continue to function even if one or more of the nodes fail.
- **Durability:** data is replicated across multiple nodes thereby reducing the risk of data loss if one or more nodes fail.
- **Performance:** multiple nodes can improve performance by allowing for parallel processing and reducing the load on any single node.
- **Security:** spreading data across multiple nodes makes it more difficult for attackers to compromise the entire data set. Data encryption and secure channels can also be employed to protect the data in transit/rest.
- **Governance:** data can be stored and processed in accordance with local policies/regulations/controls to ensure compliance.

Negatives

- **Complexity:** managing and operating distributed storage systems is complex and may require specialised technologies.
- **Consistency:** maintaining consistency across multiple nodes can be challenging, and different distributed storage systems may have different trade-offs when it comes to achieving consistency.
- **Dependency:** the reliability and performance of a distributed storage system can be affected by the reliability of the underlying infrastructure, such as the network and the individual nodes.

Option 4: distributed storage system [potential future application]

Distributed storage systems technologies

Typically you would choose distributed technology for use cases where the volume, velocity, and variety of data is high and may require near real-time processing.

Technologies for this distributed storage systems include:

- **Distributed file systems:** Distributed file systems such as Hadoop HDFS provides a way to store and manage large amounts of data across multiple servers. They allow for data to be split across multiple nodes, which improves scalability and fault tolerance.
- **Object storage systems:** Object storage systems such as Amazon S3 and Azure Blob provides a way to store and manage large amounts of unstructured data across multiple servers as objects.
- **Distributed NoSQL databases:** like MongoDB are designed to handle high-concurrency and high-availability, and they can be used for real-time data sharing.

Conclusion

Distributed storage is an important consideration for a fully distributed architecture. It offers an approach to storage of data across multiple nodes in a network in a secure, scalable, and fault tolerant way.

This offers a more attractive solution than a central database, as the workload is distributed across multiple nodes rather than being concentrated on a single node or server.

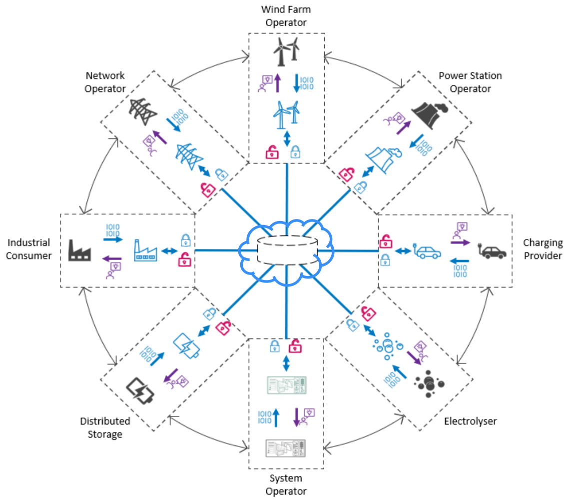
However, to fully realise a distributed architecture, other components related to distributed processing and distributed communication would also need to be considered as part of a wider technology stack.

This complete architecture will also need to consider microservices and ways to orchestrate large volumes of data being processed, stored and shared in real time, with multiple producers and consumers.

Whilst a complete distributed architecture should be the aiming point in the long term, the demonstrator use case would not require this type of storage - especially when considering the demonstrator will only involve a few network operators sharing small volumes of data.

However, this option would be suitable for future use cases where large volumes of data requires storage and consumption from multiple users. This may also include simulation and analytical purposes, or for data which is relatively static and does not require real-time updates. Furthermore, the distributed nature could require different governance and security controls on each storage node to satisfy different users.

The locality and ownership of the storage nodes remains a question; ideally, storage would reside within the producers' infrastructure, but there may use cases where the VirtualES platform would need to provide this.



Option 5: distributed streaming platform [target for use case]

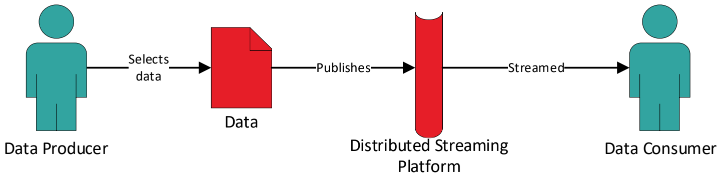
Description

A streaming platform is used to share event data between data producers (publishers) and consumers (subscribers).

The streaming platform acts as a single platform to connect multiple users to relevant data using a publisher/subscriber model.

Data is streamed using a central log by publishers, and subscribers are notified and granted access via a real-time subscription to that data.

The platform acts as a distributed log whereby data is stored as events in the order they were produced in a persistent and fault-tolerant way, thereby allowing for processing of event streams in real-time (similar to an enterprise messaging system).



Positives

- **Data retention:** data can be stored for configurable amounts of time. Once the data is processed it can be discarded as it is no longer needed, thereby minimising the amount of data requiring storage.
- **Data ownership:** consumers can retain control of their data and share the event updates with subscribers they grant access to.
- **Real-time data:** streaming platforms are designed to share real-time data/streaming data by processing the data in real-time as it is produced, rather than storing it first then processing it later, thereby ensuring that consumers ingest the latest data.
- **Schema validation:** can be used to validate certain schemas/data models before processing, therefore ensuring that only approved standards are used e.g. CIM.
- **Asynchronous data:** data production is asynchronous from consumption thereby allowing subscribers to consume data at different rates that it can control e.g. if there is a crash or systems are down for maintenance etc.
- **Parallelism:** allows for parallel distribution and consumption of data.
- **System agonistic:** a data producer/consumer can be any type of system e.g. database, cache, Hadoop system etc.
- **Performance:** low-latency and high fault-tolerance.

Negatives

- **Data storage:** streaming platforms are typically not designed to serve as a general-purpose data storage solution like a database.
- **Complexity:** streaming platform architectures are complex unless using certain technologies or frameworks like Kafka (but this is still relatively complicated).
- **Querying:** some streaming technologies are not built to perform complex queries or batch analytics - they should instead perform only simple aggregations and transactional queries.
- **Large messages:** some technologies are not designed to process large data messages.

Option 5: distributed streaming platform [target for use case]

Apache Kafka

Apache Kafka is a technology that can be used for real-time data sharing between multiple users. It is a distributed, publish-subscribe messaging system that allows for the real-time streaming of large amounts of data.

Kafka is designed for high-throughput, low-latency, and high-availability. It can handle high-concurrency and a large number of users, and can be used to share data between different systems, applications, and services.

It allows multiple producers to send data to a topic (a named data feed) and multiple consumers to read from that topic. With this publish-subscribe model, Kafka can handle a large number of concurrent reads and writes, which makes it an ideal technology for real-time data sharing. Additionally, Kafka can be integrated with other technologies, specifically open-source Apache products for analytics and machine learning.

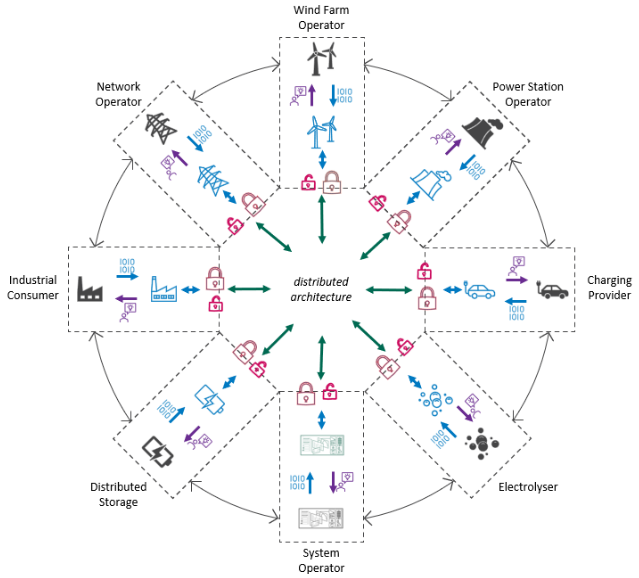
However, Kafka is not a data store in the traditional sense, it is instead designed to handle real-time data streams. However, it can be used with other data storage systems, such as a data lake or data warehouse to handle the real-time streaming of data and the batch processing of historical data.

Conclusion

A streaming technology is a logical solution for sharing of data in real-time or near real-time. Technologies such as Kafka are typically used across different industries for processing and sharing of big data in a distributed way.

A fully distributed architecture will need to adopt this solution to orchestrate the communication and distribution of data between producers and consumers. Furthermore, it can be to validate schemas/data standards to ensure that common models are shared between organisations. If coupled with distributed storage, then it can be used to coordinate the storage of data across the different databases & nodes.

Getting producers to publish their data as streams, rather than using centralised pipelines, would adhere to Data Mesh principles. The demonstrator use case could adopt a simple implementation of a streaming service to share data between the network operators in real-time and without needing to store the data, thereby avoiding the requirement for centralised ownership. Coupled with the adoption of a data portal for the user-interface and data catalogue, this could form the foundation for the VirtualES technology stack.

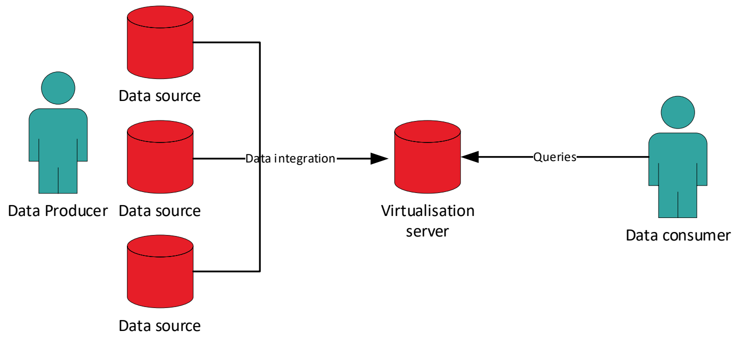


Option 6: data virtualisation [not recommended]

Description

Data virtualisation involves creating a virtual layer that sits on top of the underlying data sources and provides a unified, real-time view of the data. This virtual layer, known as a data virtualisation server, acts as an intermediary between the data sources and the consumers. This allows the data to be accessed and queried without physically copying or moving the data.

As this technique also allows data from multiple sources to be combined and presented in a single view, it eliminates the need for consumers to access each data source individually. This approach works well when organisations need a unified view of the data, but does not need to store all the data in a single location.



Positives

- **Data integration:** can simplify data integration by creating a unified view of the data, even when the data is stored in multiple systems and formats.
- **Data governance:** can improve governance by providing a centralised mechanism for defining and enforcing data policies.
- **Flexibility:** the data can be transformed and optimised for specific use cases without physically moving or copying the data.
- **Data duplication:** eliminates the need for multiple sources of the data to be stored and maintained, reducing the risk of error. This can also simplify some data security controls because you will not have to secure copies of data in different locations.
- **Real-time access:** provides real-time access to data, allowing consumers access to the latest data.
- **Abstraction:** hides the complexity of multiple endpoints.

Negatives

- **Performance:** performance can be affected for use cases involving large-scale, real-time data sharing scenarios.
- **Data quality:** does not automatically improve the quality of the data so data validation and data quality checks must be conducted to ensure the data is suitable for its desired purpose.
- **Data security:** when sharing sensitive data with multiple organisations, ensuring the security of that data can become a challenge.

Option 6: data virtualisation [not recommended]

Conclusion

Data virtualisation offers a promising solution for data integration challenges by providing a unified and virtualised view of the data from multiple sources, without moving or copying the data.

However, virtualisation architectures are complex and there is no clear requirement for the VirtualES demonstrator to provide this level of data integration.

Whilst aspects of it may be suitable for the long-term architecture, careful considerations and factors need to be considered to ensure its suitability and effectiveness when sharing large volumes of real-time data between multiple external organisations. Some key considerations pertain to the data governance and security.

Furthermore, use cases that require single views of the data from multiple disparate systems would need to drive this requirement but these currently do not exist. Considering these factors, along with its complexity, data virtualisation is not a recommended option for the VirtualES.

Conclusion

Summary of the data sharing options

Six data sharing options have been reviewed for their suitability for the demonstrator use case and the future VirtualES vision. The technology, data, security and governance considerations were used from Section 2 to derive the assessments of the options.

Considering the scale, ambition and complexity of use cases that the VirtualES may need to address in the future, any single data sharing option would not meet all the requirements for all use cases.

An amalgamation of techniques and technologies may need consideration to accommodate different use cases, volumes of data, and number of users.

Considering this, a summary of the options and their applicability for the VirtualES is provided:

- **Option 1 - point-to-point sharing:** Direct sharing (via emails) is not recommended because it does not meet the range of technology, data, governance and security requirements.
- **Option 2 - centralised datastore:** Suitable for use cases where the data has to be centralised and administered by a central owner, and where, for example, historical data is required.
- **Option 3 - data portal:** Suitable when acting as the user interface and data catalogue for users to search and discover their data.
- **Option 4 - distributed storage system:** Suitable to distribute data amongst different nodes and servers for instances where performance, security, and large volumes of data are required.
- **Option 5 - distributed streaming platform:** Suitable for sharing of real-time data, and orchestration of data sharing with users and data stores by implementing an event-driven architecture.
- **Option 6 - data virtualisation:** Is not recommended, due to the its complexity, limited current requirements, and uncertainty on the applicability of its architecture for future VirtualES.

A distributed architecture contains many complex parts; it is not necessarily just about distributed storage, but also considers distributed processing, communication, orchestration, governance, scalability etc. for data sharing with multiple users.

Key principles from Data Mesh, as described in Section 2, will also need consideration for adoption to ensure that its architecture borrows best practices for building a distributed data sharing platform, rather than becoming a monolithic big data platform with centralised ownership.

These principles not only drive the technology decisions, but also the behaviour of users, in terms of treating their data as a product (i.e. making sure it is interoperable, self-describing etc), and offering it to consumers via a serve and pull mechanism - rather than traditional push and ingest.

Based on these conclusions, the design of the demonstrator use case should adopt a technology stack which sets the foundation for further development of a future fully distributed architecture.

To illustrate this, a draft High Level Design (HLD) for the demonstrator use case, and potential options for the future VirtualES architecture are provided in Section 5.

4

Review of existing platforms

Overview & conclusion

Selection of existing platforms reviewed for the demonstrator

Background

Various organisations across the energy sector share data between external actors using different data sharing platforms. The platforms vary in architecture, type of data shared, and functionality, in order to meet bespoke requirements.

As part of the demonstrator use case, a high-level desk based study and user interviews were conducted to establish the functionality and architectural components for various data platforms.

This was completed to identify whether any existing platforms were deemed suitable for adoption as part of the demonstrator, considering the significant work which has already been undertaken in this area.

Platforms such as Skywise and Telicent CORE fall outside the energy sector but it was important to consider a broad range of existing platforms, so that we could learn from other sectors and potentially collaborate with them.

Selection of existing platforms reviewed

The following existing data sharing platforms were assessed at a high-level. It is acknowledged that this is not an exhaustive list of all platforms available.

1. DAFNI
2. GreenSync
3. Open Energy
4. NeRDA
5. Skywise
6. Telicent (CORE)

A range of existing energy sector data portals administered by the network operators were also assessed in the previously completed data portals review conducted after the VirtualES demonstrator discovery phase.

This data portal review considered factors such as the accessibility, metadata formats, number of datasets, and types of licensing.

Existing internal organisation platforms, such as the ESO Data Portal, were not reviewed at this stage.

The full data portal review report can be read here.

Conclusion of review of selected existing platforms

Most of the existing platforms demonstrate promising solutions. However, not all them necessarily meet the requirements for the VirtualES demonstrator, nor its adoption for a future fully distributed architecture.

Telicent offers a promising solution which could have a place in the future technology stack of the VirtualES, especially considering its open-source software and demonstrable value to digital twin projects.

Open Energy offers the most promising data portal solution, housing a rich curation of metadata across the energy sector using CKAN. Its Trust Framework, which is modelled on Open Banking, also offers a good solution to enable secure data sharing. Icebreaker One also expressed interest in using Open Energy as part of the demonstrator use case and combining their application with another data sharing capability – a streaming service for streaming data between users.

The combination of a streaming solution, along with Open Energy’s data catalogue and Trust Framework, and additional security and data governance controls, can offer the building blocks for the demonstrator, which will also set the foundation for future iterations of the VirtualES.

Review of existing platforms

DAFNI, GreenSync, and Open Energy

	DAFNI	GreenSync	Open Energy
Type of data (open/closed)	Closed; national infrastructure datasets	Closed; near real-time data streaming for high volume telemetry data sets	Open & closed; energy sector data
Architecture	Centralised; cloud hosted on AWS	Centralised; streaming service (based on Kafka)	Data portal; no data is stored on the platform
Data standards	Several; users can upload their own models and run workflows on them.	Several, but none preferred or required.	Several; metadata on the standards is provided
Features	Centrally managed datastore; an Extract Transform Load (ETL) framework; a data catalogue; data access and publication service; modelling service; workflow manager; visualisation suite; security service & access control	Validation & registration service (AAR); persistently accurate register (CAR), API access; minimal data storage required by streaming the data	Data catalogue with the largest curation of energy data across the sector. Users are directed to where the data is stored (with their respective owners/organisations) so that they can request access to it. It also contains a user trust mechanism for secure data sharing.
Comments	Used by CReDo. Platform has a range of sophisticated controls/features which the VirtualES may also have in the future. However, the centralised nature of the platform may not be an appropriate fit for the long-term vision for the VirtualES.	The platform uses low voltage, domestic energy data therefore it is adoption may not necessarily be suitable for the demonstrator use case	Open Energy offers a comprehensive data catalogue and Trust Framework, which is modelled on Open Banking to enable secure data sharing. Icebreaker One are also open to combine the Open Energy platform with additional data sharing capabilities (e.g. streaming data service) as part of the wider technology stack for the demonstrator use case

Review of existing platforms

Skywise, NeRDA, and Telicent (CORE)

	Skywise	Near Real-time Data Access (NeRDA)	Telicent (CORE)
Type of data (open/closed)	Closed; aviation in-flight, engineering and operational data	Closed; near real-time heat maps and energy capacity registers DNO data	Closed; defence and digital twin data
Architecture	Centralised	Data portal	Distributed streaming adopting data mesh
Data standards	Several; users can upload their own models and run workflows on them.	CSV (recognising that it will be in CIM)	Information Exchange Standard 4 (IES4) – but not limited to
Features	Different packages: Core X^1 for data discovery, dashboarding and reporting; Core X^2 add specific models, customise the ontology, create own applications; Core X^3 advanced analytics and ML	Data is stored in MongoDB and made available for consumption via an API; web based data portal.	Consumed data is cleansed, security labelled, and converted into IES4. Data can be visually represented using graphs nodes & relationships. Open and closed applications are offered as part of the tech stack; data is retained within the users' infrastructure; adopts Data Mesh principles
Comments	Types of data and intended use cases is not suitable for the demonstrator or future virtual energy system. However, features pertaining to dashboarding an reporting may be useful to consider for the future VirtualES.	Platform offers a simple solution for sharing of near real-time data between users, however, datasets shared does not encompass all the necessary data for the demonstrator and this may require changes to accommodate for CIM sharing (rather than current CSV).	Platform offers functionality and an architecture stack which largely aligns with some of the VirtualES thinking. its open source nature and involvement with digital twin projects means that it could be considered as part of the longer term VirtualES architecture and vision.

5

High Level Designs (HLD)

Overview

Summary of the HLD for the demonstrator and potential future VirtualES

Background

Based on the findings of of this report, and the requirements from EDiT, indicative HLDs have been created for both the demonstrator use case and for the future VirtualES technology vision.

The HLDs do not represent, nor contain, the comprehensive list of functional and non-functional requirements for the VirtualES. This exercise will need to be conducted at a future stage (Beta) to establish the requirements of the final solution.

These HLDs provide the outline for the approach and vision for subsequent technical architecture work to establish and agree the VirtualES technology design.

It is important to note that it was important to consider the EDiT Digital Spine concept as part of the overall architecture for the future VirtualES solution. Therefore, the concept of Digital Spine nodes are included in the HLD.

Overview of demonstrator use case HLD

The demonstrator use case HLD considers the adoption of a streaming service to publish the producer's data, and offer it to consumers in real-time, without requiring storage of the data. The shared data will be subject to appropriate security and governance controls, with access and distribution offered through secure APIs.

A data portal, which is currently expected to be Open Energy, will be adopted to offer the search, discovery and data cataloguing functionality for the use case. Furthermore, their Trust Framework could also be adopted to establish the trust protocols between users.

Overview of the future VirtualES vision HLD

The future VirtualES vision HLD considers a range of additional functionality and potential services for users. A description for each of these are provided in this section.

It is important to note that functionalities pertaining to distributed storage, centralised storage, analytical capabilities etc., are not prescribed requirements for the future of the VirtualES. Instead, they are considerations and potential options in the future but predicated on use cases and customers driving the requirements. There remains a question on the locality, ownership and governance for these services.

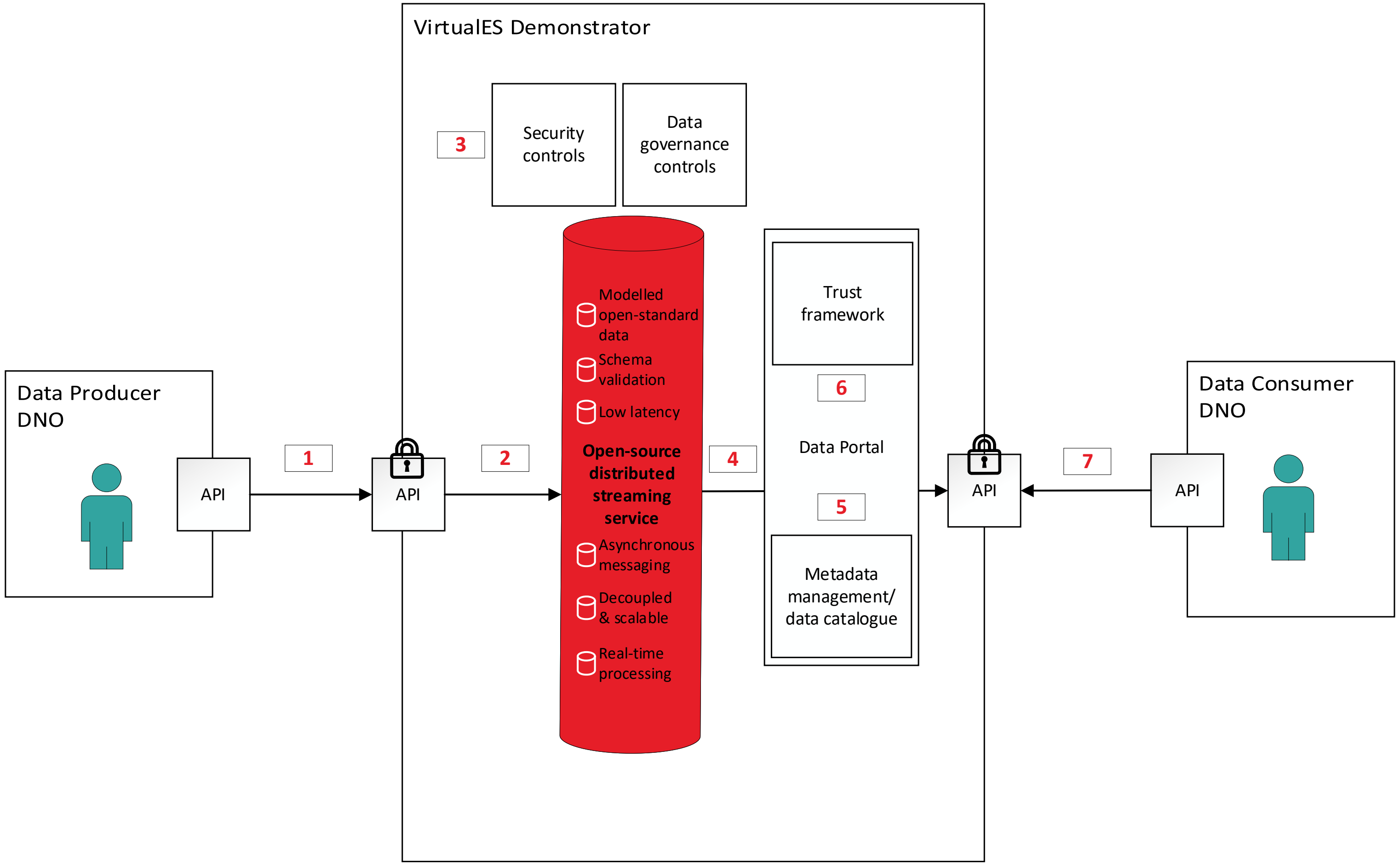
To achieve a fully distributed architecture, it is important to adopt some of industry’s best practices and emerging thinking in the realm of decentralised data sharing. Most notably Data Mesh, as previously mentioned in Section 2, which establishes a set of principles to achieve effective data sharing in a distributed way.

These principles not only have implications for the technology aspects, but also to the people and processes, more specifically how data producers treat their data so that it adheres to *data as a product* thinking i.e., making sure the data is readable, appropriate data quality etc., before it is offered (preferably as a stream of data) to consumers.

This will mean that data owners should have ownership of their data, and are responsible for treating their data accordingly so that it is sufficiently fit for purpose when sharing with consumers using the VirtualES.

Demonstrator use case HLD

Proposed HLD for the demonstrator



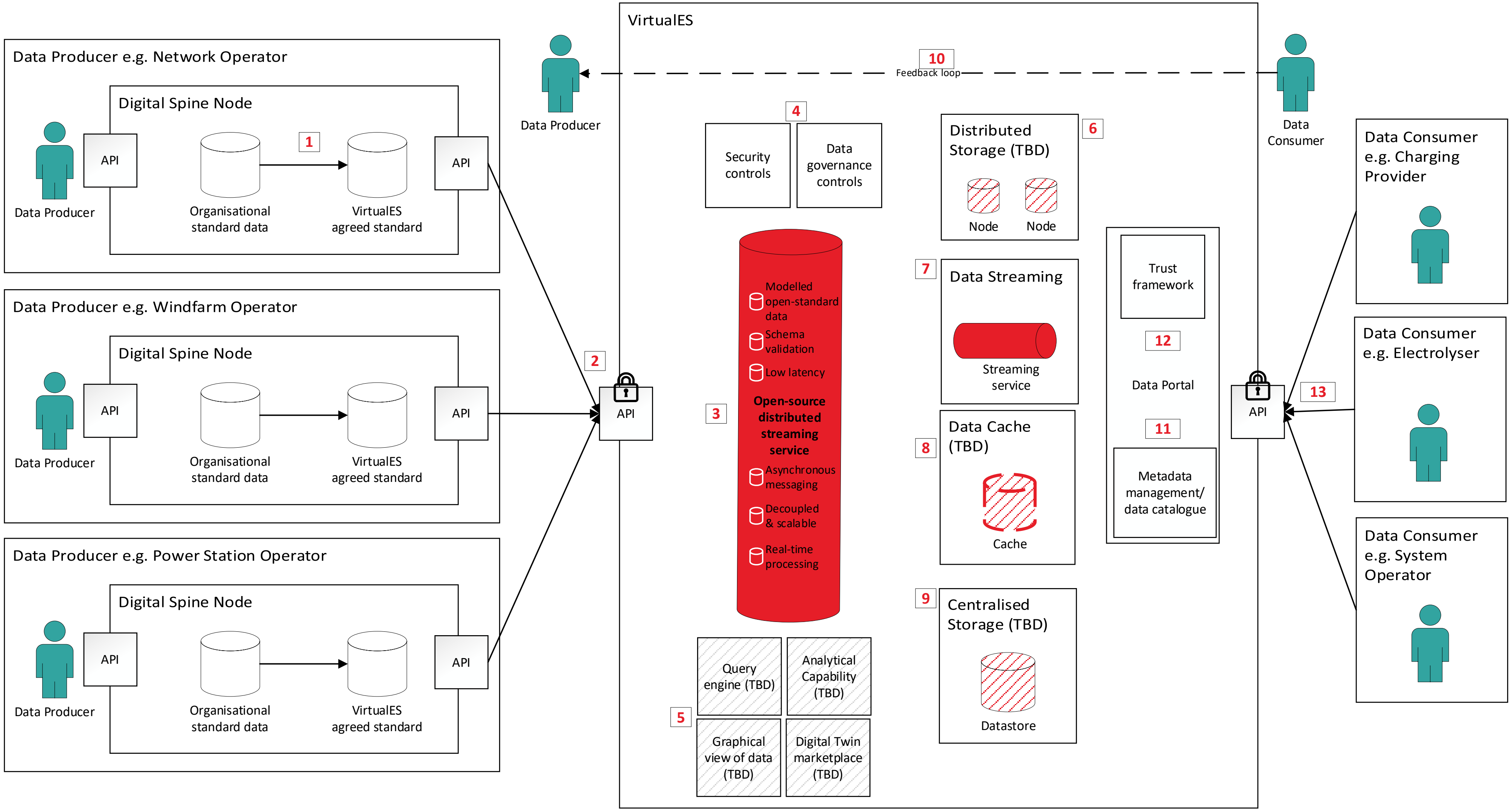
Description of diagram

The description numbers below align with the labels on the diagram.

1. Data producer indexes and publishes (including permissions tags) required data via secure APIs.
2. Data is published to a distributed streaming service, where the data is validated to ensure approved schemas and standards are used e.g. CIM.
3. Published data has security and data governance controls implemented, along with user permissions as part of the Trust Framework (see 6).
4. Data is streamed in real-time with consumers and does not require storage within the demonstrator.
5. Data portal i.e. Open Energy, allows for data to be searchable and discoverable. Consumers can request access to the required data using the data catalogue.
6. Trust framework is used to apply trust controls between participants, as part of Open Energy.
7. Data consumers consume required data via secure APIs, after they have searched for their data, requested it, and undergone governance and security protocols.

Future VirtualES vision HLD

Proposed HLD for the future VirtualES



Data Producer e.g. Network Operator

Digital Spine Node

Data Producer

API

Organisational standard data

1

VirtualES agreed standard

API

Data Producer e.g. Windfarm Operator

Digital Spine Node

Data Producer

API

Organisational standard data

VirtualES agreed standard

API

Data Producer e.g. Power Station Operator

Digital Spine Node

Data Producer

API

Organisational standard data

VirtualES agreed standard

API

Data Producer

2

API

VirtualES

10

Feedback loop

4

Security controls

Data governance controls

3

Modelled open-standard data

Schema validation

Low latency

Open-source distributed streaming service

Asynchronous messaging

Decoupled & scalable

Real-time processing

5

Query engine (TBD)

Analytical Capability (TBD)

Graphical view of data (TBD)

Digital Twin marketplace (TBD)

Distributed Storage (TBD)

6

Node

Node

7

Data Streaming

Streaming service

8

Data Cache (TBD)

Cache

9

Centralised Storage (TBD)

Datastore

Trust framework

12

Data Portal

11

Metadata management/ data catalogue

Data Consumer

API

13

Data Consumer e.g. Charging Provider

Data Consumer e.g. Electrolyser

Data Consumer e.g. System Operator

Future VirtualES vision HLD

Proposed HLD for the future VirtualES

Description of diagram

The description numbers below align with the labels on the diagram on the previous page.

1. **Digital Spine node:** offers a mechanism to transform data producer's organisational standard data to an agreed and interoperable standard for sharing with the VirtualES.
2. **Data producers:** publish required data via secure APIs to the VirtualES. There may be defined data contracts in place between the consumers and producers which sets out the expectations and agreements for the shared data e.g. data quality, readability, security controls or labels etc.
3. **Distributed streaming service:** data is validated to ensure approved schemas and standards are used e.g. CIM. The streaming service is also used to orchestrate the processing, storage and sharing of large volumes of data using event-driven architecture patterns.
4. **Published data:** has security, data governance and metadata management controls implemented.
5. A range of optional capabilities/applications are available for users. This includes:
   - **Query engine:** allowing for querying of data across the VirtualES.
   - **Analytical capability:** analytical capability, which may include ML use cases.
   - **Graphical view of data:** for graphical representations of data using nodes & relationships, or other visualisation techniques.
   - **Digital twin marketplace:** service allowing users to consume digital twin models.
6. **Distributed storage:** for performance, security, storage and availability requirements. This may be used for analytical and simulation use cases, and also to enable storage of large datasets or metadata which may need historical and aggregated views. The datastores may reside within the organisation's premises, or within the VirtualES depending on the use cases which will drive the requirements. A streaming service can be used in conjunction to publish changes to the databases, where the changes are consumed in real-time e.g. Change Data Capture.
7. **Data streaming:** data is streamed in real-time with consumers and may not require storage within the VirtualES.
8. **Data cache:** cached data for performance requirements when sharing large files. This may be a semantic view of the data without a need to access the underlying data.
9. **Centralised storage:** for data that requires limited distribution. This may be for sensitive data, or for instances where data needs to be stored for analytical purposes e.g. data warehousing. The datastore may reside within the organisations' premise, or within the VirtualES depending on the use cases which will drive the requirements.
10. **Feedback loops:** feedback mechanism between the consumers and producers for their data and digital twin models.
11. **Data portal:** allows for data to be searchable and discoverable. Consumers can request access to the required data using the portal. There may be instances where the portal can also offer web views of rendered data e.g. CIM.
12. **Trust framework:** is used to apply trust controls, as proposed by Open Energy.
13. **Data consumers:** consume required data via secure APIs, after they have searched for their data and requested access.

6

Recommendation

Recommendations

Overall recommendations that are applicable to the demonstrator

Based on the findings of this report, the following recommendations are given:

- A fully distributed architecture for the VirtualES is envisioned. This will likely involve a combination of technological solutions to accommodate multiple future use cases, and to meet a range of technology, data, and security & governance requirements. Furthermore, the design should adopt Data Mesh principles to deliver an architecture suited for decentralised data sharing.
- The technology stack for the demonstrator use case will need to both fulfil the requirements of the demonstrator, and also set the foundation for future developmental iterations of the VirtualES to build towards a fully distributed architecture, as illustrated in the future HLD.
- Open Energy is considered a suitable data portal for the demonstrator use case, due to its rich collection of metadata, cataloguing capability, and Trust Framework. There is an opportunity to collaborate to deliver a technology stack which incorporates streaming data, along with various security and governance controls for the VirtualES demonstrator.
- The recommended HLD for the demonstrator technology (and likely host) will comprise of:
  - A **data portal** for cataloguing and discovery of the data.
  - A **distributed streaming service** for sharing of data in near real-time whilst avoiding storage of the data.
  - The validation of a common data standard (e.g. CIM) by the streaming service, to enable interoperability.
  - Governance, security and user trust mechanism controls, which includes secure access via APIs for publishing and consumption of data.

ARUP

Virtual Energy System

Powered by **ESO**

VirtualES@nationalgrideso.com