Policy-Aware Federated Query Orchestration Across Energy Data Spaces and Edge AI Services

Vasile Ofrim
TU Cluj-Napoca, Romania
vasile.ofrim@cs.utcluj.ro

Mihai Daian
TU Cluj-Napoca, Romania
mihai.daian@cs.utcluj.ro

Dragos Lazea
TU Cluj-Napoca, Romania
dragos.lazea@cs.utcluj.ro

Liana Toderean
TU Cluj-Napoca, Romania
liana.toderean@cs.utcluj.ro

Anca Hangan
TU Cluj-Napoca, Romania
anca.hangan@cs.utcluj.ro

Tudor Cioara
TU Cluj-Napoca, Romania
tudor.cioara@cs.utcluj.ro

ABSTRACT

The execution of AI driven analytical tasks across distributed data sources and edge services remains challenging due to heterogeneous governance policies, computational constraints, and volatile edge environments. In this demonstration, we present a modular federated orchestration framework for governance-aware query processing in smart grids. The system maintains a continuously updated context model of energy assets, computational resources, and participating entities, enabling policy-aware planning. Natural language operator intents are translated into executable federated workflows using large language models. A custom energy data space connector enforces local edge data policies while supporting secure multi-node data exchanges and associated computation. A federated orchestrator coordinates distributed workflow execution, ensuring policy compliance, infrastructure feasibility, and adaptive re-optimization under dynamic conditions.

PVLDB Reference Format:
Vasile Ofrim, Mihai Daian, Dragos Lazea, Liana Toderean, Anca Hangan, and Tudor Cioara. Policy-Aware Federated Query Orchestration Across Energy Data Spaces and Edge AI Services. PVLDB, 19(12): 4654 - 4657, 2026.
doi:10.14778/3827998.3828089

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://gitlab.com/policy-aware-federated-query-orchestration.

1 INTRODUCTION

Data spaces are a promising paradigm for enabling secure and sovereign data sharing across organizational boundaries. In the energy domain, they aim to facilitate collaboration between multiple stakeholders, including utilities, distribution system operators (DSOs), service providers, energy communities, and prosumers, while preserving control over data access and usage policies. At the same time, Edge AI services are increasingly deployed closer to decentralized energy infrastructure to support real-time analytics and operational decision-making, such as energy grid monitoring, demand forecasting, and energy resources optimization. However, executing queries that combine distributed data sources and Edge AI services across these stakeholders remains challenging, as orchestration mechanisms must operate across decentralized infrastructures while respecting heterogeneous governance and policy constraints.

In emerging energy data space environments, stakeholders such as DSOs often lack the technical expertise required for data management systems. They typically express their analytical needs as high-level intents or natural language requests rather than formal queries. This intent-driven approach introduces additional orchestration challenges, as systems must translate natural language inputs into executable federated queries that account for distributed data sources and Edge AI services [4, 6]. Furthermore, query planning must respect data access policies, computational constraints, and energy infrastructure requirements, all while operating in volatile environments where edge resources may become temporarily unavailable [8]. Current state-of-the-art federated query processing systems exhibit limitations in this regard. They primarily focus on data integration and performance optimization [7], offering limited support for enforcing data space policies and governance constraints during query orchestration [2, 3, 5]. They also assume relatively stable data sources and expect users to formulate queries using formal query languages [1].

This paper addresses these limitations through a modular orchestration framework for querying and integrating distributed data sources and Edge AI services in smart grids, shown in Figure 1. The framework maintains a unified, continuously updated representation of the computational and energy infrastructure, including grid topology, computational resources, and participating organizations, enabling context-aware query orchestration. Natural language operator queries are automatically translated into executable federated workflows, with large language models bridging the gap between human intent and technical execution. Interactions with edge nodes are mediated through custom data space connectors that enforce local-only data policies while supporting multi-node computations, enabling secure publication, discovery, negotiation, and transfer of data under contractual constraints. A federated orchestrator coordinates workflow execution across distributed edge nodes, ensuring policy compliance, feasibility under infrastructure and computational constraints, and resilience to volatile edge resources. Together, these components enable federated query analytics across heterogeneous energy data spaces.

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3828089

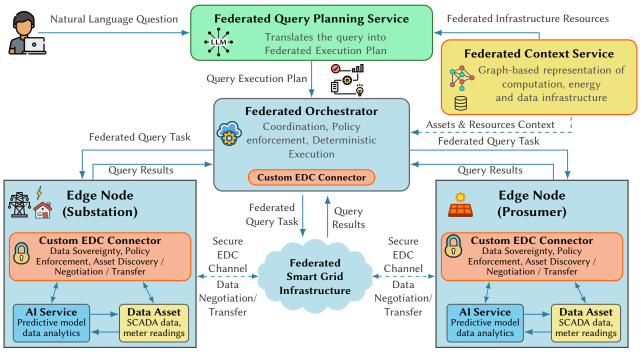

Figure 1: Architecture overview

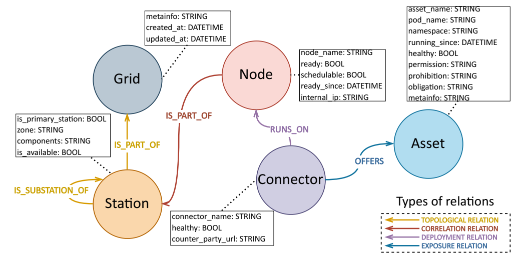

Figure 2: Context state graph data model

2 CORE ARCHITECTURAL COMPONENTS

2.1 Federated Context Service

The Federated Context Service maintains a unified, continuously updated representation of a distributed smart grid infrastructure covering energy grid assets, computational resources, and energy dataspace participants. It serves as the metadata backbone for policy-aware orchestration and federated query planning across edge environments.

In smart grid environments, where data and computation are inherently distributed, physical assets such as stations, substations, and distributed energy resources define the topology constraints. The edge computing nodes host AI-driven workloads, and energy dataspace connectors expose governed assets under contractual policies. As these layers evolve independently, they will produce dynamic context state updates. Therefore, we define a unified, graph-based metadata abstraction that correlates heterogeneous infrastructure information, providing support for policy-compliant federated planning, asset discovery, and edge workload orchestration. This unified data model (see Figure 2) represents the infrastructure as a property graph in which *Stations* belong to a *Grid* and may follow a hierarchical structure through `IS_SUBSTATION_OF` references. Each *Station* contains one or more *Nodes*, describing computing hosts with attributes related to their state and network connectivity. *Connectors* running on *Nodes* define inter-node communication endpoints and expose *Assets*, associated with deployed workload instances, described by pod-level identifiers, namespace, health status and usage policies. The model enables cross-layer traversals in a single query, by co-locating topological, deployment, and exposure relations in a single graph. The structure is designed to generalize to any cloud-edge deployment governed by data sovereignty policies. Building on this representation, the service ensures separation of concerns and event-driven consistency through deterministic updates of the graph structure, while incorporating governance policies as first-class metadata.

The service continuously synchronizes the context state from three sources. Static energy grid infrastructure (stations, substations, topology) is ingested from structured configuration artifacts and defines the physical backbone of the graph. Cluster-level metadata is ingested from the Kubernetes control plane via event-driven watchers running on independent threads to decouple event listening from graph persistence and preserving low-latency capture under high throughput. For each node, the service tracks: identity and network metadata, readiness and scheduling state, temporal state and grid position. Dataspace connectors are discovered as specialized workloads. Upon lifecycle events, the service updates connector metadata (identity and hosting node), retrieves and synchronizes exposed data asset catalogs, persists asset metadata and associated access control policies. The AI driven workloads (e.g., energy forecasting) are correlated with dataspace assets while their deployment and health state are continuously tracked and reflected in the context state graph.

2.2 Federated Query Planning Service

The Federated Query Planning Service translates the operator queries expressed in natural language into validated, executable plans for the federated orchestrator. Its primary role is to interpret, plan and validate workflow plans that the orchestrator can map onto the available infrastructure and execute. Therefore, it bridges human operator intent and automated orchestration in a policy and data governance aware manner.

The service begins with the interpretation stage, where the operator's natural language question is transformed into a structured JSON representation. A semantic prompt is constructed using the question, session context, and allowed assets and substations. The model generates multiple candidate interpretations, which are scored and selected based on contract compliance, confidence, and temporal correctness. The selected candidate is then normalized, filtered against the asset catalog, and subjected to strict contract checks, ensuring correctness before planning. The service follows a model-driven planning approach that emphasizes control and interpretability over purely deterministic pattern matching. For interpretation and and planning, it uses two specialized models, Qwen2.5-Instruct (1.5B and 3B), fine-tuned for the smart grid domain deployed on small compute nodes. The interpretation model produces contract-compliant JSON structures capturing relevant resources and temporal context, while the planning model transforms these structures into a validated, step-by-step workflow plan. Each step links assets, nodes, substations, and workflow dependencies into an executable schema.

A human-in-the-loop clarification mechanism is defined to ensure accurate interpretation of operator queries. When the interpretation model generates a candidate plan with insufficient confidence (e.g. ambiguous asset references, missing temporal information, etc.), the service prompts the operator for clarification. During this loop, the specific ambiguities or contract violations are identified in the JSON candidate. It generates targeted clarification questions or suggestions that the operator can answer to resolve uncertainty (e.g., "Which substation should this asset be assigned to?" or "Specify the desired time window for the analysis step."). The operator's responses are incorporated, and the interpretation is re-validated against the constraints. In this way the JSON interpretation is both semantically correct and operationally feasible, while preserving automation for most queries. Also, operators maintain control over critical planning decisions, due to human oversight. Finally, session memory is updated, and the fully executable plan payload is returned to the orchestrator for execution. Each step is normalized with unique IDs, explicit outputs, and verified HTTP endpoints.

2.3 Custom Energy Data Space Connector

The Custom Eclipse Dataspace Components Connector mediates all interactions between edge nodes and distributed AI workloads, ensuring that local data exchange policies are enforced. It acts as a gateway between edge participants, enabling data asset publication, discovery, negotiation, and transfer. The connector implements a provider-consumer model for secure data exchange. Data providers register their assets and associate them with asset-level policies that define catalog visibility, discovery and access permissions, negotiation rules, and locality restrictions governing where an asset may be processed. Contracts bind the data assets to policies, determining who can discover and access them. The control plane manages asset publication, catalog exposure, policy evaluation, and contract negotiation, the process through which consumers acquire access rights. A data consumer retrieves the provider catalog, negotiates access, and receives a data plane endpoint and access token for asset retrieval. The data plane handles the actual transfer of negotiated data assets between consumer and provider.

On top of the standard connector functionalities, we defined two extensions to support federated edge analytics: a Data Plane References Resolver and a Response Cache Storage. They allow the orchestrator to validate and compose workloads deterministically, supporting federated query execution and policy-aware orchestration without exposing raw asset data. Edge deployments depend on locality-constrained data assets that must remain within their originating node or station, while still serving as inputs to distributed AI workloads and federated services. The Data Plane References Resolver addresses this challenge by enabling clients to include URI references instead of raw data in service requests. During invocation, the references in the payload are intercepted. Data references are resolved locally at the node level, and the federated request is rewritten with the retrieved values prior to execution. This mechanism allows AI workloads to consume local assets transparently, without exposing sensitive data across nodes and orchestrated computations.

The orchestrator may validate asset operability during federate execution without accessing full or sensitive result data. Response Cache Storage extension provides interception of requests marked for caching via HTTP headers, temporary in-memory storage of responses, with a link generated for limited reuse, where expiration is coordinated to release memory once the intended accesses are completed. This approach reduces unnecessary data transfers and preserves local data sovereignty.

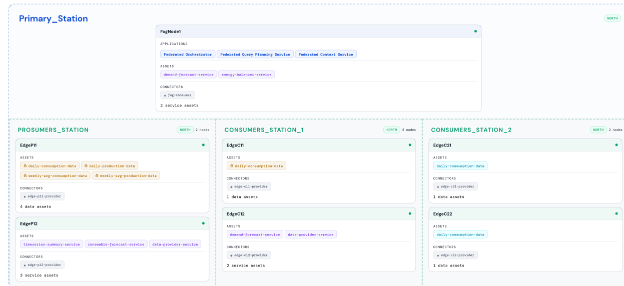

Figure 3: Demo computational, data and energy grid infrastructure

2.4 Federated Orchestrator

The Federated Orchestrator is the central coordination layer for secure, policy-compliant, deterministic execution of federated query plan across distributed edge nodes. Execution begins with infrastructure context discovery, querying the Federated Context Service to obtain grid topology, node status, data space connectors, and registered assets. The orchestrator validates the plan against the current infrastructure state, adjusts assignments as needed, and maps abstract workflow steps to concrete nodes and connectors for execution. Node availability and data asset presence are checked. If an edge node becomes unresponsive, it adapts the plan by shifting the execution step to an alternative node or rejects the plan if no accessible asset can be found. The data space connectors negotiation pipeline ensures secure, authorized access to assets, embedding governance awareness directly into the execution process.

Finally, the orchestrator executes the plan step-by-step, resolving inter-step references so that outputs from earlier steps become inputs to subsequent ones. It invokes services on edge nodes and negotiates secure data access, enforcing governance and contractual policies. Service invocation on edge nodes leverages either streaming references for local-only or large data, or cached responses to minimize redundant transfers, performing data-efficient and edge-native operations. This design enforces edge-native execution and data-efficient operations, as steps are executed close to data sources and intermediate outputs can be cached or streamed to avoid unnecessary transfers. In the final step, the orchestrator streams the output back to the client.

3 DEMONSTRATION USE CASES

We consider a distributed smart grid environment composed of four stations: a Primary Station, a Prosumer Station, and two Consumer Substations (see Figure 3). Each station hosts one or more nodes that manage energy telemetry and run analytics. The infrastructure combines edge computing resources deployed close to energy assets with fog/cloud nodes providing coordination and optimization capabilities. Consumer Substation 1 hosts edge nodes C11 and C12, and Consumer Substation 2 hosts a single edge node C21. These nodes maintain local consumption measurements and participate in federated query execution through exposed analytical services. Each consumer edge node hosts a Demand Baseline Service that computes local consumption averages from historical data. The

Figure 4: Energy system imbalance and federated query plan

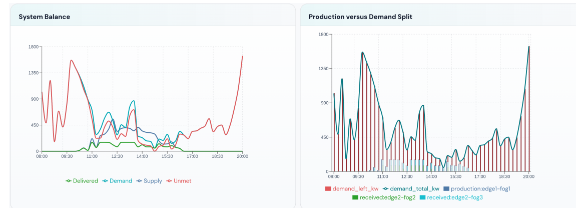

Figure 5: Energy system balancing, production and demand allocations

Prosumer Station comprises two edge nodes, P11 and P12, which manage both consumption and local renewable generation data. These datasets remain under prosumer control and serve as inputs to the local demand and renewable forecasting. Node P12 supports controlled data access within the station. The Primary Station hosts system-wide coordination services, including an Energy Balancer Service that aggregates derived metrics from distributed nodes and evaluates redistribution feasibility subject to grid constraints.

To reflect realistic smart grid conditions, the system enforces data governance policies that restrict the movement of raw sensor data. Datasets such as daily consumption time series and historical consumption records cannot leave the station where they are generated. Instead, computations are executed locally, and only derived results or aggregated outputs are shared during federated query execution. Latency measurements show an average of 2s for intent interpretation (1.5B model) and 44s for plan generation (3B model), resulting in about 46s to generate an 11-step plan on an 8GB GPU. A clarification round adds only 2s. Across 30 runs, latency varied by less than 0.2%.

Use Case 1: The DSO query **"Can the smart grid re-distribute excess solar energy from prosumers to nearby consumers today?"**

Although the operator's question appears simple, answering it requires governance-aware federated query execution. Raw consumption time series and historical datasets are subject to prosumer and consumer station-level locality constraints and cannot be transferred to other stations or to the cloud. As a result, surplus detection cannot be performed centrally. In our approach the planner generates a federated execution plan, and the orchestrator coordinates distributed computations across stations, returning only aggregated or derived results (see Figure 4). From a federated analytical task perspective, the query is decomposed into locality-aware subtasks. For each prosumer node, the weekly average energy consumption is computed using locally stored historical data that cannot leave the edge node due to governance constraints. Current solar production and consumption metrics are then retrieved at the same edge node. Nodes where real-time production exceeds the weekly baseline are identified, and the available surplus energy is estimated locally. The orchestrator discovers nearby consumer nodes and invokes Demand Forecasting Services to evaluate whether the surplus can be redistributed without exceeding grid transmission capacity. Only derived metrics (e.g., surplus values and demand forecasts) are forwarded to the Energy Balancer Service, which performs the final feasibility assessment.

Use Case 2: The DSO query **"The solar production has dropped unexpectedly due to rapidly changing weather conditions. Can the redistribution plan still be executed?"**

The stochastic nature of renewable generation introduces continuous updates to production time series at prosumer edge nodes. When actual solar output deviates significantly from forecasted values, previously computed surplus estimates become invalid, potentially violating the energy redistribution constraints. Rather than re-executing the entire federated planning task, the orchestrator performs incremental re-evaluation of the distributed query plan. Therefore, only affected prosumer nodes recompute their local surplus using updated production data, preserving governance constraints that do not allow for raw data movement. The recomputed surplus values are propagated as updates for the grid distributed state to the federated orchestrator. The orchestrator performs partial re-evaluation of the Energy Balancer Service, updating global plan decisions without triggering full re-calculation across all stations (Figure 5). Afterwards the downstream allocations are revised accordingly.

ACKNOWLEDGMENTS

This research was funded by the European Union's Horizon Europe programme (Grant Agreement No. 101136216, Hedge-IoT, https://hedgeiot.eu/).

REFERENCES

[1] Nikos Fotiou, Chalima Dimitra Nassar Kyriakidou, Athanasia Maria Papathanasiou, Vasilios Siris, and George Polyzos. 2026. Relationship-based Access Control for Data Spaces. *Data Science and Engineering* (2026), 1–17. https://doi.org/10.1007/s41019-025-00320-y
[2] Victor Giannakouris and Immanuel Trummer. 2025. Rethinking Pluggable Federated Query Optimization: From Laptops to Data Warehouses. *Proceedings of the VLDB Endowment* 2150 (2025).
[3] Anhelina Kovach, Leticia Montalvillo, Jorge Lanza, Pablo Sotres, and Aitor Urbieta. 2026. Understanding data spaces: A Systematic Mapping Study of foundations, technical building blocks, and sectoral adoption. *Computer Science Review* 59 (2026), 100819. https://doi.org/10.1016/j.cosrev.2025.100819
[4] Andrea Manzi, Raul Bardaji, Ivan Rodero, Germán Moltó, Sandro Fiore, Isabel Campos, Donatello Elia, Francesco Sarandrea, A Paul Millar, Daniele Spiga, et al. 2025. interTwin: Advancing Scientific Digital Twins through AI, Federated Computing and Data. *Future Generation Computer Systems* (2025), 108312. https://doi.org/10.1016/J.FUTURE.2025.108312
[5] Christoph Quix. 2025. Toward Advanced Query Processing in Dataspaces. In *DOLAP*. 88–90.
[6] Zijie Su, Muhammed Tawfiqul Islam, Mohammad Goudarzi, and Adel N Toosi. 2026. LLM-Driven Intent-Based Privacy-Aware Orchestration Across the Cloud-Edge Continuum. *arXiv preprint arXiv:2602.16100* (2026). https://doi.org/10.48550/arXiv.2602.16100
[7] Jiayi Wang and Guoliang Li. 2025. Aop: Automated and interactive llm pipeline orchestration for answering complex queries. In *Conference on Innovative Data Systems Research (CIDR)*. https://vldb.org/cidrdb/papers/2025/p32-wang.pdf
[8] Siyuan Xia, Zhiru Zhu, Chris Zhu, Jinjin Zhao, Kyle Chard, Aaron J. Elmore, Ian Foster, Michael Franklin, Sanjay Krishnan, and Raul Castro Fernandez. 2022. Data station: delegated, trustworthy, and auditable computation to enable data-sharing consortia with a data escrow. *Proc. VLDB Endow.* 15, 11 (July 2022), 3172–3185. https://doi.org/10.14778/3551793.3551861