Reflections from our experimentation in supporting sensor data in the Floods and Droughts Research Infrastructure (FDRI) programme

A deliverable report for the AMPLIFY-EDS project

Authors:

Gordon Blair [1] https://orcid.org/0000-0001-6212-1906

Faiza Samreen [1] https://orcid.org/0000-0002-9522-0713

Mike Brown [1] https://orcid.org/0000-0002-2996-0633

Philip Trembath [1] https://orcid.org/0000-0002-3690-2941

Matt Fry [1] https://orcid.org/0000-0003-1142-4039

Dominic Ginger [1]

Hollie Cooper [1] https://orcid.org/0000-0002-1382-3407

Helen Rawsthorne [1] https://orcid.org/0000-0002-6540-8547

Nathan Shaw [1]

Affiliations:

1 - UK Centre for Ecology and Hydrology

Date: 26 November 2025

Contents

Context

We believe this report will be useful by providing an important case study - the Flood and Droughts Research Infrastructure (FDRI) - which will both inform the design of the sensor data commons at the heart of AMPLIFY-EDS and provide feedback on specific design choices around technologies and semantic constructs.

About the Environmental Data Service

The Environmental Data Service (EDS) provides a focal point for scientific data and information spanning all environmental science domains: atmosphere and climate, earth observation, polar and cryosphere, marine, terrestrial and freshwater, geoscience, solar and space physics. The EDS is made up of a network of distributed data centres, with domain specific expertise.

Our main goal is to ensure that environmental data are made available, accessible and re-usable for the long-term in order to fully realise their value. We are funded by the Natural Environment Research Council (NERC), part of UK Research and Innovation, to advise researchers on how to prepare data for long term storage and dissemination.

The EDS is a fundamental part of NERC's digital strategy and works with the Digital Solutions Hub, and other partners, to break down disciplinary barriers and facilitate data sharing beyond academic use.

About the project

The AMPLIFY-EDS project was funded by the UKRI DRI Phase II call. It ran from March 2024 to Oct 2025.

Within this project, the EDS developed an end-to-end sensor workflow to demonstrate elements of an EDS data commons framework - initiating the first elements in the data commons roadmap. It delivered live data from several research sensor sources through a common workflow and demonstrated their use within shared tools.

AMPLIFY had two primary aims:

- To co-design and develop a live-data prototype service that delivers environmental sensor data and its associated metadata in a standardised, end-to-end workflow.
- To engage with peer sensor data initiatives and stakeholders across the UKRI Digital Research Infrastructure (DRI) and beyond, ensuring alignment, interoperability, and community input.

This work builds upon previous work undertaken via the ENHANCE and BOOST projects.

Introduction

The overall goal of Activity 1.1 is to develop a sensor data commons for the NERC Environmental Data Service (EDS), looking at the overall design of the necessary EDS enhancements to support sensor data and the full end-to-end pathway from acquisition to ingestion and subsequent analyses.

As part of this, UKCEH contributed to supporting reflections and activities (formally, sub-task 1.1B) designed to support the architectural work underpinning the sensor data commons. Specifically, we focussed on the core underlying sensor infrastructure underpinning the Floods and Droughts Research Infrastructure (FDRI). This £38 million project is establishing a nationwide Floods and Drought Research Infrastructure, offering near real time data to the hydrological community. Led by UKCEH, the project is deploying instruments for observing our water environment – measuring evaporation, soil moisture, weather, groundwater and river flows. The overall components of FDRI are captured in figure 1.

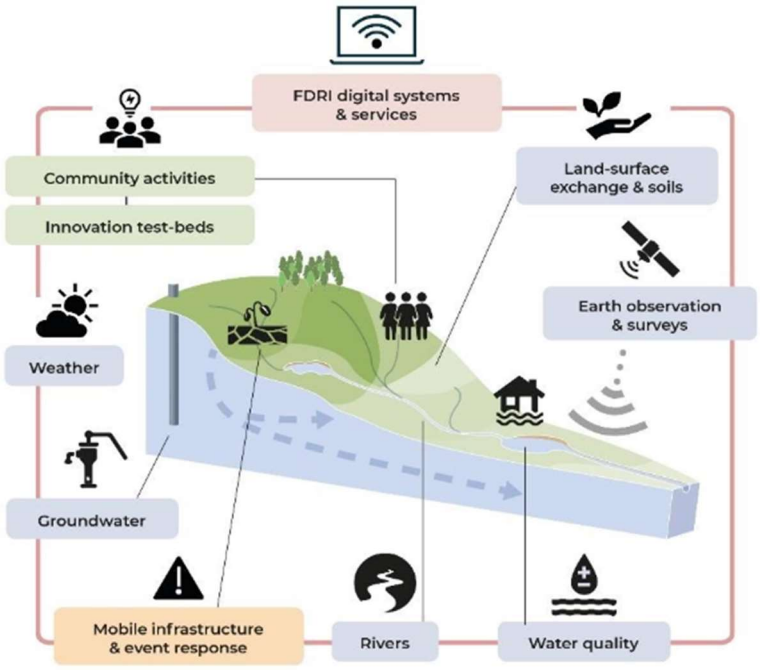

Figure 1: Illustration of the key components of the FDRI infrastructure within one FDRI river basin

There is a very significant digital element of this project (circa £10m) devoted to developing the necessary Digital Research Infrastructure (DRI) to support FDRI ambitions, with a strong focus on the pathway from sensor infrastructure to the ingestion of quality assured data into the EDS. FDRI therefore provides a strong case study to both inform the design of the sensor data commons at the heart of AMPLIFY-EDS, and provide feedback on specific design choices around technologies and semantic constructs. This short document presents the results of this work.

Digital research infrastructure for sensor data: pilot implementation

Over the last 18 months, we have been developing a pilot Digital Research Infrastructure for FDRI targeting initially streaming, time-series data. To provide further focus, we initially focussed on targeting the COSMOS-UK soil moisture network, as indicative of the issues of managing high-volume streaming data from an extensive monitoring network. The COSMOS-UK network provides near-real time soil moisture data from across the UK for use in a variety of applications including farming, water resources, flood forecasting and land-surface modelling.

The pilot supports the end-to-end pipeline from data acquisition to analysis based on this network. The overall architecture is shown in figure 2.

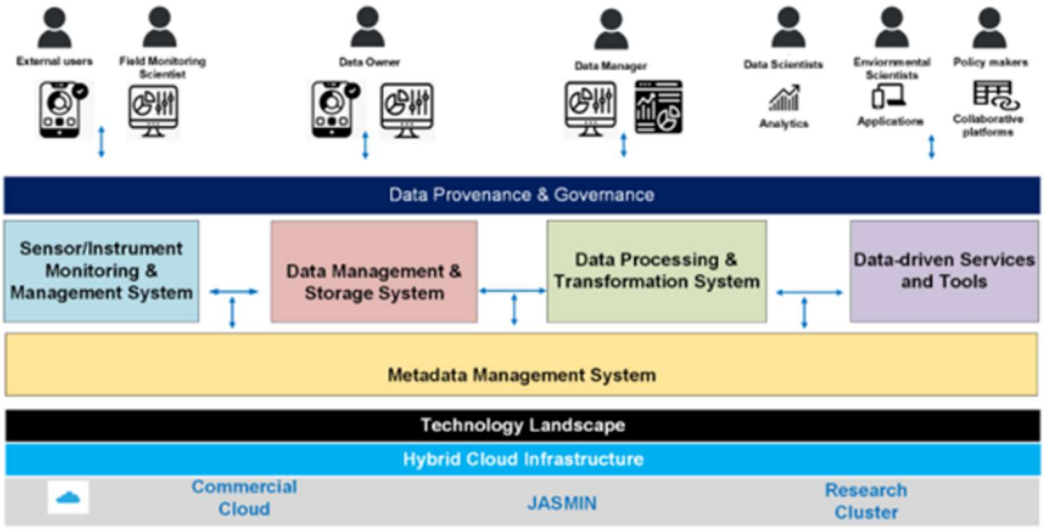

Figure 2: FDRI digital components

Built on cloud technologies and cloud-native principles, this infrastructure ensures portability, scalability and network independence, enabling us to meet emerging requirements and use cases such as real time or near real time data streaming across diverse data types including time series, images and more.

Key features include:

- **End-to-end pipeline** that ingests raw sensor data, processes it in real-time, and presents it through dashboards and APIs.
- **Network-agnostic architecture** supports multiple monitoring networks (COSMOS, NRFA).
- **Metadata driven** configuration and processing rules, enabling quick sensor onboarding without code changes.
- **Cloud-native infrastructure** deployed on Amazon Web Services.

A more detailed view of the streaming architecture is provided by figure 3.

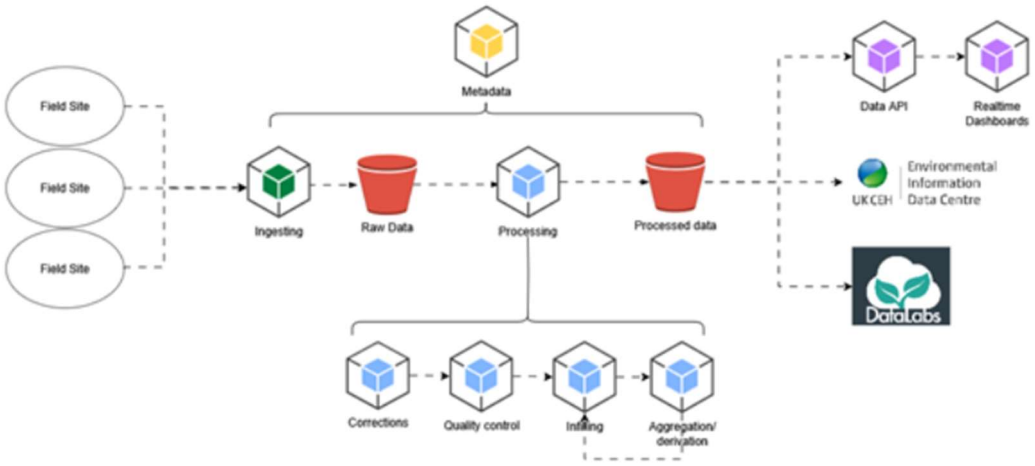

Figure 3: The FDRI software pipeline for streaming time-series data

We are using AWS IoT Core to manage live data streams, capable of scaling up to 1 million data streams per second. We store data at various levels to capture provenance and ensure data quality control. To keep everything modular, we encapsulate functionality into distinct components. For managing data processing workflows, we leverage Argo Workflows. All of this is running on a Kubernetes cluster hosted on AWS, with separate production and staging environments for better isolation. Infrastructure monitoring is performed using Prometheus and Grafana, ensuring continuous health checks and identifying areas for potential improvements over time.

While this has been prototyped for COSMOS-UK data, we are now in the process of evaluating the generality of the approach for other sensor data sources (EA, SEPA, NRW).

Workshops and events

As stated above, this experimental work has provided input into AMPLIFY-EDS and the design of the sensor data commons. This interaction has been managed by three key events as listed below:

- Workshop 1: FDRI / AMPLIFY meeting on sensor data models (webinar: 24/1/25)
- Workshop 2: Deep dive into technical architectures (webinar: 15/5/25)
- Workshop 3: Discussion with NEON (Discussion: 27/08/25)

Further details of these events can be found in appendix A.

Overall findings and recommendations

1. The architecture shows what can be achieved in terms of developing Digital Research Infrastructure on top on the robust and flexible service architectures offered by contemporary cloud platforms, using recognised and/or emerging standards;

2. Where possible it is important to be platform independent, achieved by adhering largely to cloud native principles across the architecture;

3. The architecture has achieved our overarching goal of supporting the end-to-end pipeline from data acquisition from sensors through quality assurance to the potential ingestion into data repositories (with some potentially high impact work still to be done to fully automate the ingestion process);

4. It is crucial to integrate both the processing of data and the (automated) collation of meta-data to support this ingestion process, ensuring the resultant data elements are well described and fully discoverable (this achieving the desired bringing together of different elements of the project as targeted in the final 6 months of the project);

5. Workshop 1 was important for the project demonstrating that there was a substantial level of consensus across the consortium around both the services and standards to be used in the construction of software pipelines as well as with the semantic architectures to describe the services;

6. Overall, our experiences indicate that architectures like this can be transformative for the community in providing real-time or near-real time access to environmental monitoring data;

7. Significant effort was put into designing a pipeline that is generic and potentially applicable to a wide range of sensor types and modalities, enhance reuse as generic DRI within UKCEH, across the NERC EDS family, and across other organisations (currently being tested in FDRI and beyond as mentioned above);

8. While the OGC Sensor Things API (STA) is the best fitting open standard for the FDRI data some concerns were identified regarding its data model limitations, implementation complexity, and usability, which led us to adopt a dual strategy of offering a partial STA implementation in parallel with a simple REST API [https://nora.nerc.ac.uk/id/eprint/540576/].

Appendix A: Workshop details: agendas, attendees and discussion elements

Workshop 1: FDRI / AMPLIFY meeting on sensor data models

Date: 24/01/2025	Time: 10:30 – 12:00	Location: Online (Teams)

Agenda

10:30	Welcome
10:35	AMPLIFY-EDS brief overview
10:40	Forth-ERA brief overview
10:45	FDRI brief overview
10:50	FDRI sensor context
10:55	Epimorphics data model
11:20	Questions and discussion
11:50	Next steps and close

Attendees

Name	**Affiliation**	**Comments** (e.g., role, data interests)
Hollie Cooper	UKCEH	Hosting, data manager
Matthew Fry	UKCEH	FDRI digital lead
Helen Peat	BAS	
Mike Crosier	BAS	
Petra Ten Hoopen	BAS	
Paul Breen	BAS	
Alex Tate	BAS	
Alice Fremand	BAS	
Emma Bee	BGS	
Andrew Kingdon	BGS	
Carl Watson	BGS	
Edd Lewis	BGS	
Martin Nayembil	BGS	
Khalil Ahmed	Epimorphics	

Name	Affiliation	Comments (e.g., role, data interests)
Dave Reynolds	Epimorphics	CTO, Linked Data, ontologies and architecture for production services
Tom Guilbert	Epimorphics	
Martin Juckes	NCAS	
Philip James	Newcastle University	
Helen Snaith	NOC	
Lousie Darroch	NOC	
Alexandra Kokkinaki	NOC	
Sam Pepler	STFC	CEDA Curation manager and AMPLIFY-EDS PI
Graham Parton	STFC/CEDA/NCAS	NCAS Data Activity Lead; CEDA-Atmospheres co-lead. Working on NCAS instrument data pipelines. (also PIDS for Instruments, Complex citations and connections with BOOST-EDS and AMPLIFY-EDS)
Richard Kingston	Univ. of Manchester	
Peter Hunter	University of Stirling	
Kathryn Harrison	UKCEH	
Gordon Blair	UKCEH	Head of Environmental Digital Strategy
Mike Brown	UKCEH	Technical Lead for FDRI Digital
Faiza Samreen	UKCEH	Software Systems Architect
Mollie Cooper	UKCEH	
Philip Trembath	UKCEH	
Richard Smith	UKCEH	Software dev/data scientist, working on FDRI
Simon Stanley	UKCEH	
Rod Scott	UKCEH	
Dominic Ginger	UKCEH	
Helen Rawsthorne	UKCEH	Data Scientist – Semantic Specialist

Discussion

Alex (questions to Kal Ahmed on data model:

NOCare working with DCAT (CDIF profile of DCAT)– good to align models.

NOC plan to have a profiles of the SensorThingsAPI so that EDS can discuss variables that will be using, etc. KA: Yes this is needed, happy to join in.

On I-ADOPT – noticed included the unit but usually this is stored closer to measurement than variable. KA: being looked at.

And in the model of deployment model, a deployment can have events, e.g. start, end, calibration…. KA: this is being based on PROV so there are relations between activities…

Alex Tate: Could the 'Fault' concept be extended out to a more general sensor 'Event' of which fault is a subtype?

Are we going to have a SPARQL endpoint to bring graphs together. KA: currently building a stack that uses a triple store as the back end and layers a simple REST API on the top to provide developer friendly access and define different views to this. In long term could also have a SPARQL endpoint potentially. Have used Apache Jena / Fuseki for this.

And finally think about performance…

Could apply to have a PID for sensors (this would need integrating with the systems for acquisition and management of sensors / other assets)

Lou Darroch: have just set up an initial PID service… working with German group (https://pid-sms-tst.bodc.uk/). Would be good to continue discussions around SensorThings. KA: one of the issues with SensorThings being that dataset is defined at the sensor level rather than the variable/time series level

Sam Peplar: sounds all too harmonious! Where do there seem to be any issues where things might grate or be incompatible.

AK: SensorThings great on the sensor data point of view but for discovery this needs the DCAT level info as well. EOSC group discussing these things was describing things on the same lines…

Lou: where is uncertainty being captured… MF: in terms of content expect to have

Carl Watson: have been working with SensorThings API and have an aging viewer of this. We should work together on developing this stuff together. Phil James has some things up already in his tech stack.

Peter Hunter: what about mobile / moving sensors? KA: possible at the activity level and interested to see how to bring this into the model.

AK: For observable properties, how are we planning to manage them. Phil Trembath: keen to use NVS to manage these terms but we need to be flexible and need to understand how we can draft terms in NVS .

Other

Ideas for further meetings (and please note if you think these are already happening or if it would be helpful for FDRI to convene):

- Technical discussion on pipelines and software stacks
- I-ADOPT, NVS and observableProperties
- Compatibility of sensor info with NOC work on sensor metadata and PIDs

Workshop 2: Deep dive into technical architectures

Date: 15/05/2025	Time: 10:00 – 11:30	Location: Online (Teams)

Agenda

10:00	Intro to the session: Mike Brown
10:05	Quick Round Table
10:10	FDRI Infrastructure (inc Q & A): Faiza Samreen and Mike Brown
10:30	FDRI Metadata management and sensor Data API (inc Q & A): Dave Kal
10:50	Software testbed infrastructure based on Apache Airflow and NiFi (inc Q & A): Phil James
11:05	Forth-ERA developments and roadmap (inc Q & A): Peter Hunter

Round up, and suggestions for next meeting - all

Overarching questions to frame presentations and discussion

- What key technologies, platforms, and frameworks did you choose for building your sensor management system, and what influenced those decisions?
- How is your infrastructure designed to support sensor data ingestion, storage, and processing at scale (e.g., cloud, on-prem, hybrid)?
- How do you handle metadata management and ensure data provenance, traceability, and integrity across your sensor data lifecycle?
- What strategies, frameworks, or standards do you use to expose sensor data via APIs?
- Any lessons learned regarding cost, performance, system complexity and technology stacks?

Attendees

Participants from BGS, BODC, BAS, UKCEH, NOC, Newcastle University, University of Stirling

Discussion

During the discussion, various technology stacks were explored, and lessons were shared around the complexities of managing large sensor networks and ensuring scalability using different orchestration engines such as Argo Workflows, Apache NiFi, and Apache Airflow. A common theme was the emphasis on open-source, scalable tools to meet future requirements, along with the importance of open communication protocols for interoperability.

A cloud-native approach was central to all solutions, applied at different levels, focusing strongly on sustainability. This included the use of Kubernetes clusters for containerised orchestration and leveraging cloud platforms such as AWS and Microsoft Azure for their managed services and scalability advantages.

The discussion also raised important questions about data formats and data architectures needed to handle diverse datasets. This led us to reach out to NEON to learn from their extensive experience in data management and sharing within a mature ecosystem (see below).

Workshop 3: Discussion with NEON

Date: 27/08/2025	Time: 16:00 – 17:00	Location: Online (Teams)

Agenda

The first discussion was cantered on understanding their approach to handling time series data and also around publishing and working with users to ensure the data are FAIR. and was followed by email exchanges addressing key questions as outlines below. The meeting was organised around the following questions:

1. A high-level overview of your current data architecture or any insights? (Particularly how you have structured it to support both internal use and open public access.)

2. Beyond object storage, are you leveraging other services or data stores (e.g., file systems, databases, data lakes, streaming platforms)?
    - Any particular roles they play in your pipeline?
1. What data formats have you found most effective for streaming, querying, and long-term archiving?
    - Any lessons learned around schema evolution or format performance (e.g., Parquet vs Avro vs JSON)?
2. How do you manage data versioning and consistency across your services, especially for public datasets?
5. Any strategies have you used to balance cost and performance in your data services (e.g., tiered storage, caching, serverless vs reserved compute)?
6. Do you use any access controls, quotas, or caching layers to optimise and manage usage from external users?
7. Have you developed internal policies or monitoring tools to handle excessive access patterns to open data?

Summary of discussion

1. High-level insights into data architecture on GCP

Question: How is the GCP data architecture structured to support both internal use and open public access?
Response:

- Everything is built using **infrastructure automation (Terraform)** to ensure repeatability and manageability.
- **Role impersonation** is used for secure and flexible resource creation.
- **Projects are separated** by function (e.g., long-term storage, general storage, databases) for clearer billing and governance.
- **Data products are distributed via Cloud Storage**, leveraging **Autoclass/tiering** for cost optimization.
- Automated **disaster recovery (DR)** is configured for critical buckets.

2. Use of other data services beyond object storage

Question: What other services or data stores are leveraged, and how do they fit into the pipeline?
Response:

- Uses a wide range of GCP services: **Cloud SQL (Postgres), GKE, BigQuery, Stackdriver (logging/metrics/tracing), Cloud Run, Cloud Functions, Pub/Sub**, IAM integrations, **Artifact Registry**, and **Secrets Manager**.

- Some tools are **self-hosted on GKE** for flexibility and cost reasons:
    - **Apache Airflow** (vs. Cloud Composer) → lower cost, more DNS control.
    - **Apache Kafka** (via Strimzi) → supports offline/backfill data and avoids vendor lock-in.
    - **VictoriaMetrics** → Prometheus-compatible, cheaper, supports backfill; over a trillion datapoints managed efficiently.
    - **Grafana Loki** for edge log aggregation (cloud aggregation planned).
    - **Traefik** as load balancer; **Trino** as legacy warehouse (being phased out).

3. Data formats for streaming, querying, and archiving

Question: Which formats are most effective for different data types and use cases?
Response:

- **Streaming:** Apache **Avro**, moving to **Protobuf** for next-gen loggers.
- **Metadata streaming: JSON** (CloudEvents schema).
- **Warehouse & analytics:**
    - Legacy → **ORC** (optimized for Presto/Trino).
    - Current → **Parquet** (dictionary encoding, compression via ZSTD).
- **Specialized data:**
    - **HDF5** (to be replaced by Parquet/GeoTIFF), **LAZ/LAS** for LIDAR.
- Data now stored **by day and by sensor**, avoiding complex joins — "**no query is faster than any query**."

4. Lessons on schema evolution and format performance

Question: What lessons have been learned regarding schema evolution or file format performance?
Response:

- **Avro:** Good for schema management but file nonce breaks hash tracking; lacks unsigned types.
- **Parquet:** Fast with Apache Arrow, not appendable but efficient for columnar compression.
- **Schema management:** Avoid breaking changes; new fields are optional. Using **schema registry** for Avro; manual tracking for Parquet.
- Moving toward **Protobuf + gRPC** for new loggers; evaluating **Buf** for schema governance.
- **JSON:** Loosely used; CloudEvents standard works well for event signaling.
- Majority of data still distributed as **pre-baked CSVs**—Parquet migration will enable dynamic CSV generation.

5. Balancing cost and performance

Question: How do you manage cost vs. performance trade-offs?
Response:

- Heavy use of **Autoclass tiering** for cost-effective storage.
- **Spot instances** in GKE provide up to **10x savings**; pipelines (Airflow, Argo) can restart on reclaim.
- **Reserved compute** (1–3 years) used once workloads stabilize.
- Fine-tuned **pod resource requests** to maximize packing efficiency.
- **Workload placement optimization:**
    - Serverless (Cloud Run) → for infrequent tasks.
    - Kubernetes → for frequent workloads.
- Minimal persistent VMs (<5 total).
- Limited CDN use; **performance acceptable without caching**.

6. Access control, quotas, and usage optimization

Question: How are external access, quotas, and usage managed?
Response:

- **Public vs. private separation** via distinct GCP projects.
- Private data requires **credential-based access**; **Requestor Pays** model under consideration.
- Actively pursuing **Internet2 egress waivers**.
- **Alerts and monitoring** used for cost anomalies; no hard quota limits yet.

7. Policies and monitoring for open data access

Question: What internal measures exist to handle excessive access or misuse?
Response:

- **Automated GCP spend alerts** by project.
- No incidents requiring throttling so far — high access would be a "good problem to have."

Overall Takeaways:

- **Automation-first, cloud-native architecture** with strong separation of concerns.
- **Self-hosted open-source tools** chosen where cost, control, or backfill support matter.
- **Parquet + Kafka** emerging as the backbone for scalable, query-free data access.

- **Sustainability, cost efficiency, and vendor neutrality** guide architectural decisions.
- Future focus: **Protobuf schemas, dynamic CSV generation, and expanded data aggregation/monitoring.**