Spatial data processing with workflow engines

Pirmin Kalberer
@implgeo
Sourcepole AG, Zurich
www.sourcepole.com

Spacial data processing

- **Typical tasks**
  - Data updates / imports
  - Data exports / publication
  - Raster processing
  - Tile cache updates
- **Classic tools**
  - Scripts (GDAL CLI, Python, ...)
- **Orchestration**
  - Cron Jobs
  - Makefiles
  - Custom Scripts
  - DB controled processes

Challenges

- Maintenance (many scripts, different backends)
- Execution overview (many jobs)
- Distributed processing (multiple servers)
- Containerization (dependencies)
- Log aggregation and analysis
- Process Monitoring with alerting (failures)

Data science → Tools!

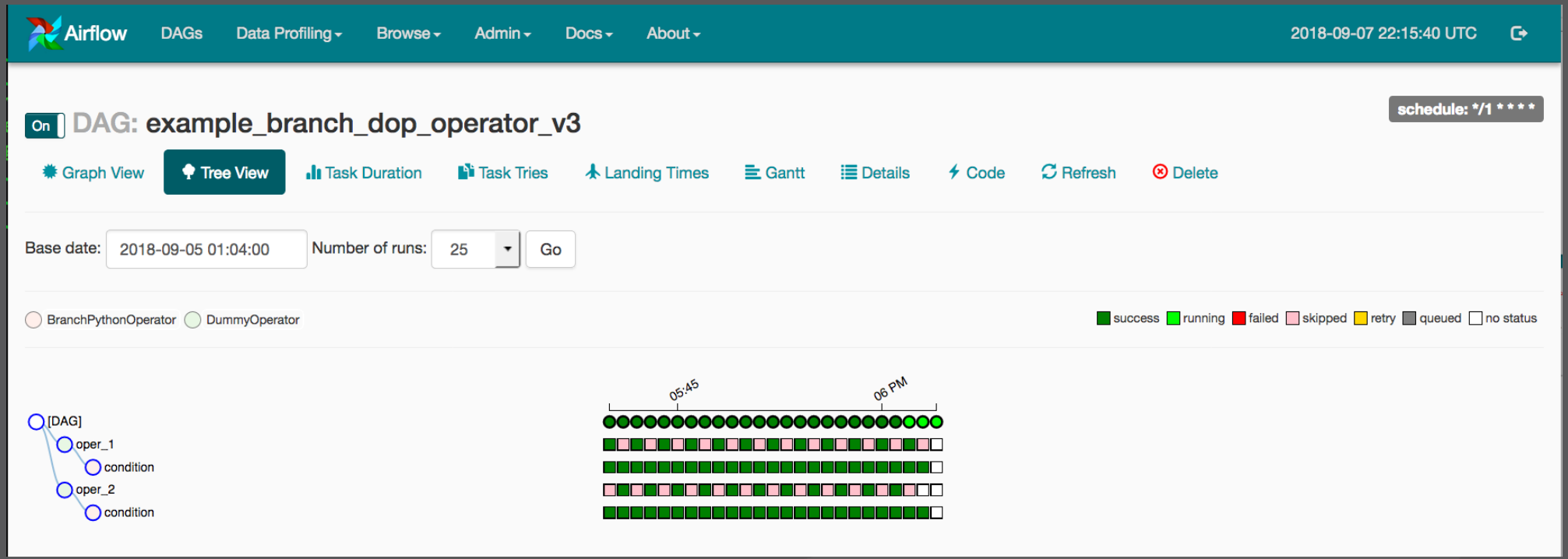
awesome-workflow-engines

A curated list of awesome open source workflow engines

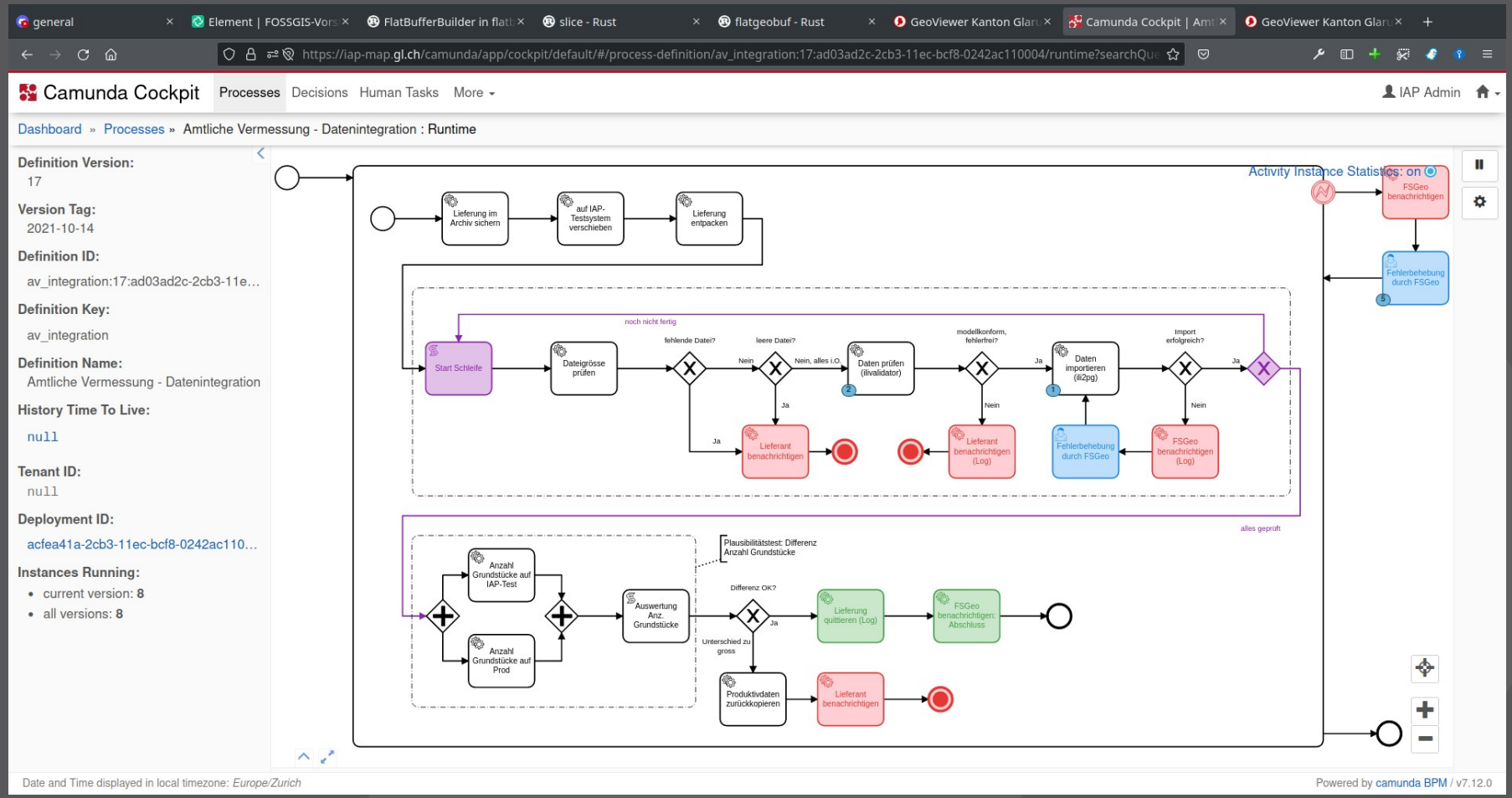
Full fledged product

- Airflow stars 27k - Python-based platform for running directed acyclic graphs (DAGs) of tasks
- Argo Workflows stars 12k - Open source container-native workflow engine for getting work done on Kubernetes
- Azkaban stars 4.1k - Batch workflow job scheduler created at LinkedIn to run Hadoop jobs.
- Brigade stars 2.4k - Brigade is a tool for running scriptable, automated tasks in the cloud — as part of your Kubernetes cluster.
- CabloyJS stars 633 - A Node.js full-stack framework with workflow engine, based on koa + egg + vue + framework7.
- Cadence stars 6.2k - An orchestration engine to execute asynchronous long-running business logic developed by Uber Engineering.
- Camunda stars 2.8k - BPMN-based workflow engine that can be embedded as java library (e.g. Spring Boot) or used standalone, including a graphical modeler and operations tooling.
- CGraph stars 328 - A simple-used and cross-platform DAG framework based on C++17 without any 3rd-party.
- CloudSlang stars 220 - Workflow engine to automate your DevOps use cases.
- Conductor stars 4.7k - Netflix's Conductor is an orchestration engine that runs in the cloud.
- Copper stars 220 - A high performance Java workflow engine.
- Couler stars 754 - Unified interface for constructing and managing workflows on different workflow engines, such as Argo Workflows, Tekton Pipelines, and Apache Airflow.
- Covalent stars 137 - Workflow orchestration platform for quantum and high performance computing.
- Cromwell stars 828 - Workflow engine written in Scala and designed for simplicity and scalability. Executes workflows written in WDL or CWL

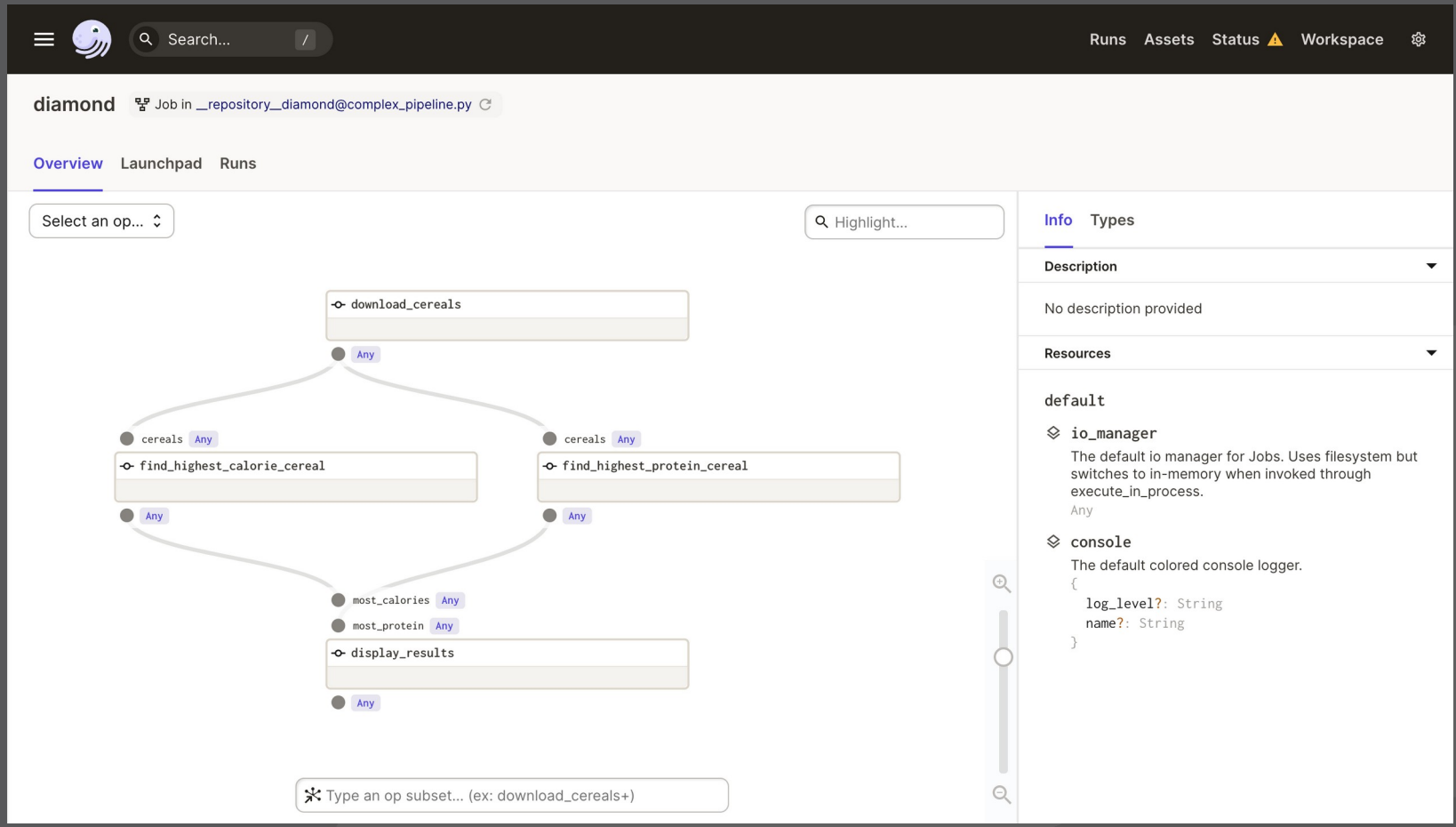
Apache Airflow – DAG (directed acyclic graph)



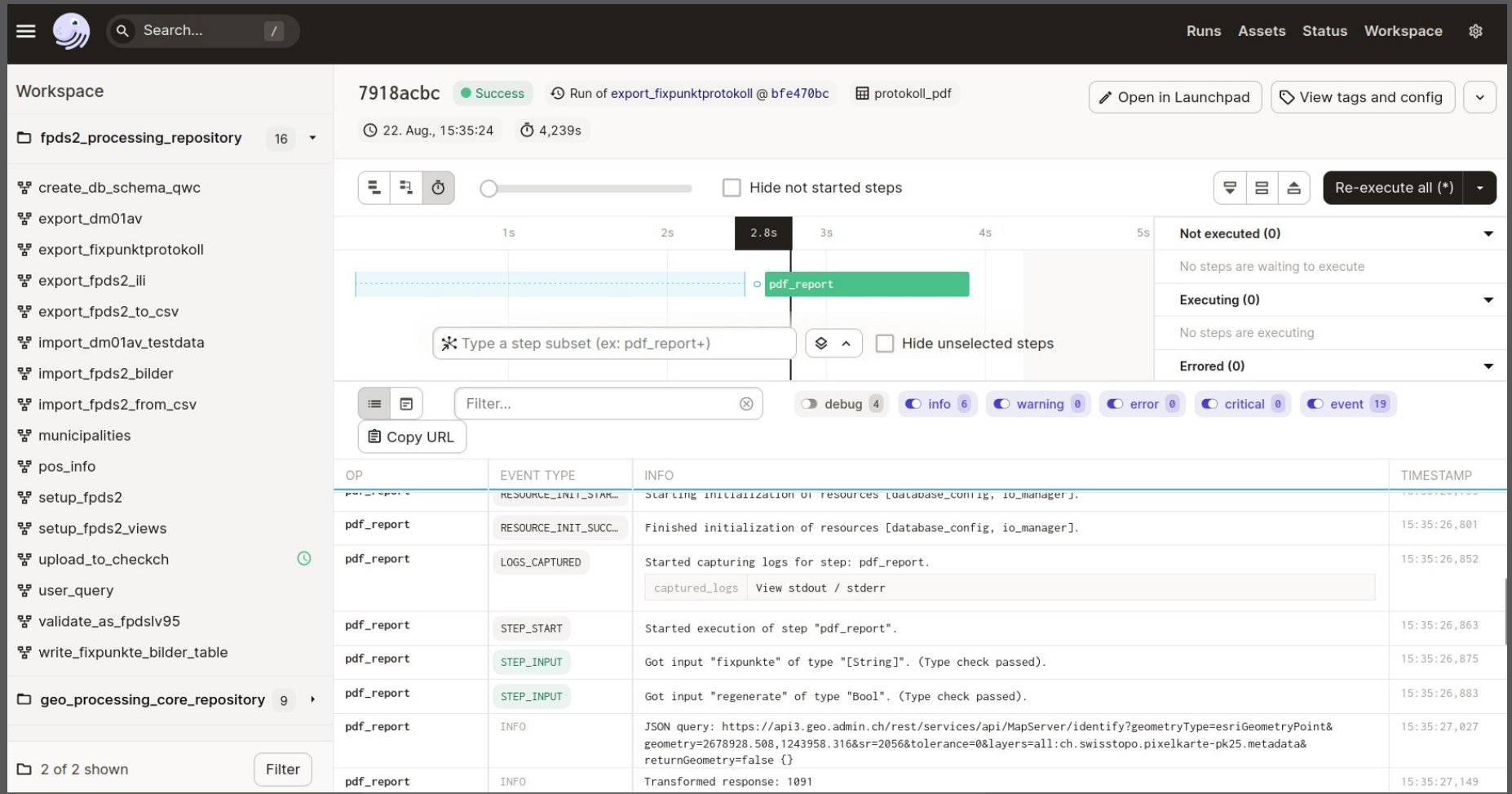
Camunda - BPMN-based workflow engine



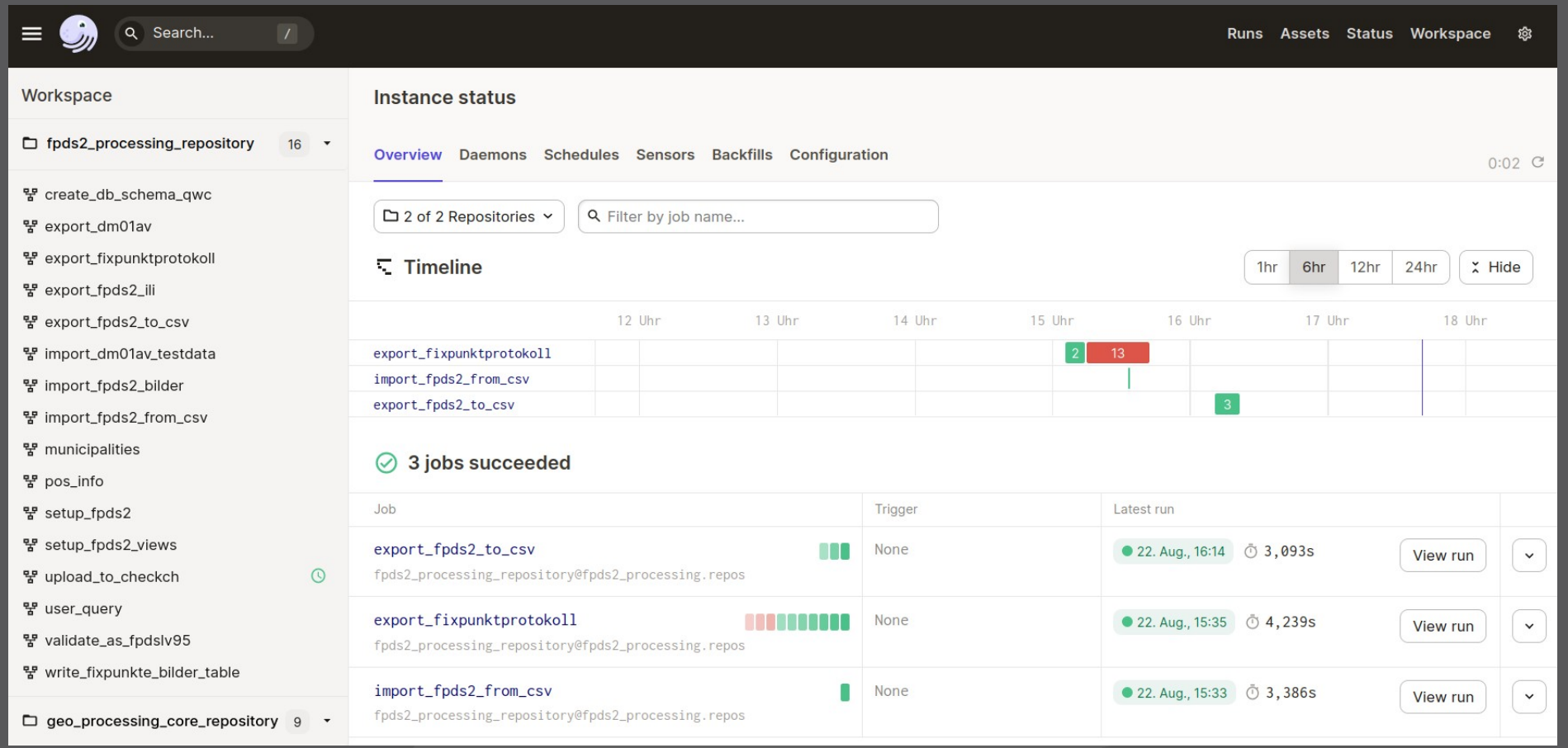
Dagster – mutli backend processing engine



Dagster – log collection



Dagster – execution overview



Dagster backend libraries

- **Processing libraries / backends**
  - Pandas
  - dbt
  - Spark
- **Runtime environments**
  - Python
  - Celery, Dask
  - Docker, Kubernetes
- **API**
  - Python
  - GraphQL

OGC API - Processes

- https://ogcapi.ogc.org/processes/

OGC API - PROCESSES

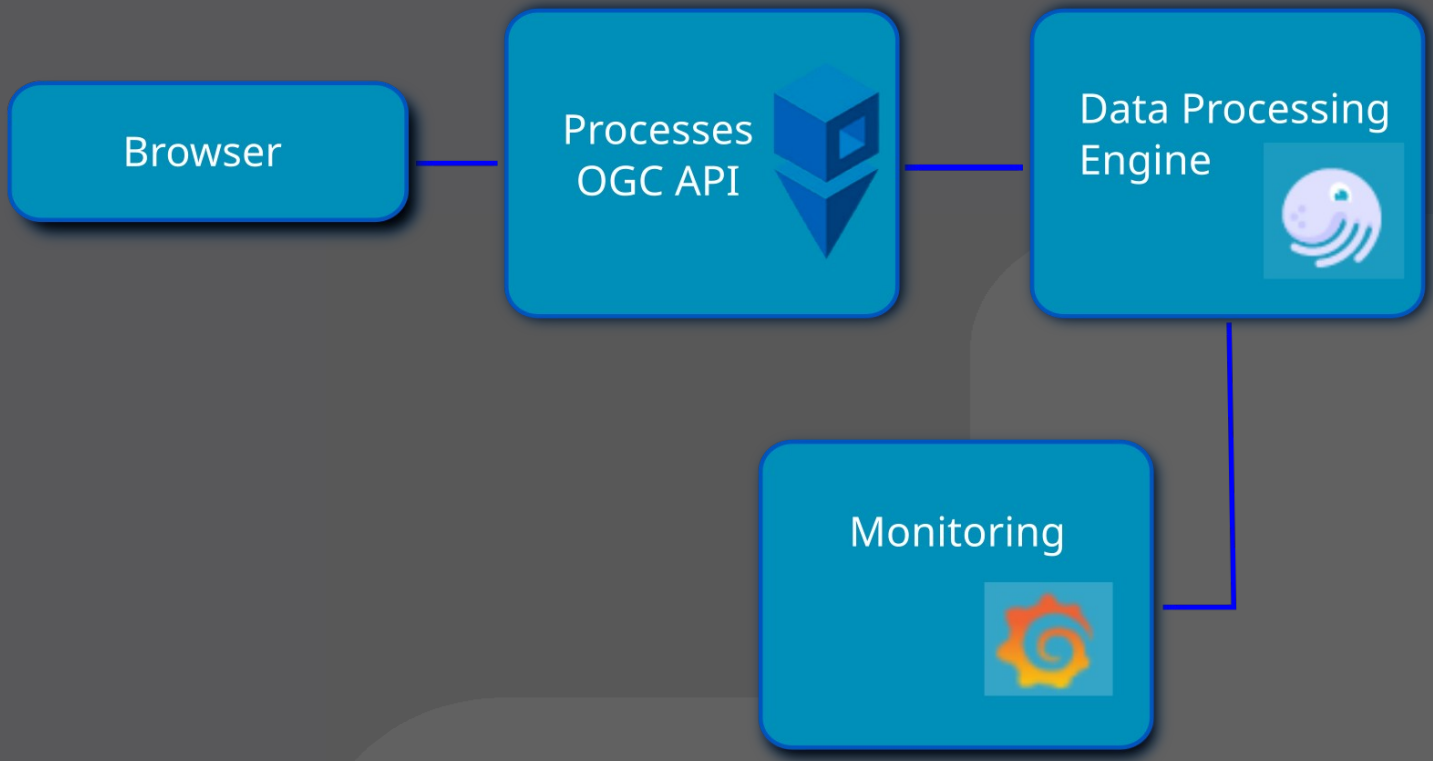
The OGC API - Processes standard supports the wrapping of computational tasks into executable processes that can be offered by a server through a Web API and be invoked by a client application. The standard specifies a processing interface to communicate over a RESTful protocol using JavaScript Object Notation (JSON) encodings. Typically, these processes execute well-defined algorithms that ingest vector and/or coverage data to produce new datasets.

Overview

- Visit the overview page to learn more about this API.

Developer-friendly OpenAPI definitions

Processing engine + OGC API + Monitoring



Thank you!

Pirmin Kalberer
@implgeo