Geodatenverarbeitung mit Workflow-Engines

Pirmin Kalberer
@implgeo
Sourcepole AG, Zürich
www.sourcepole.ch

Prozessierung Geodaten

- **Typische Vorgänge**
  - Datenaktualisierung / -import
  - Datenexport / -publikaton
  - Raster Prozessierung
  - Tile Cache Update
- **Herkömmliche Tools**
  - Skripte (GDAL CLI, Python, …)
- **Herkömmliche Orchestrierung**
  - Cron Jobs
  - Makefiles
  - Spezifische Skripte
  - DB-gesteuerte Prozesse

Problemstellungen

- Wartbarkeit (verteilte Skripte, unterschiedliche Backends)
- Übersicht über Ausführung (viele Jobs)
- Verteilte Prozessierung (mehrere Server)
- Komplexe Abhängigkeiten (Ausführung in Container)
- Log Aggregation und Analyse
- Prozess Monitoring mit Alarmierung

Data science → Tools!

awesome-workflow-engines

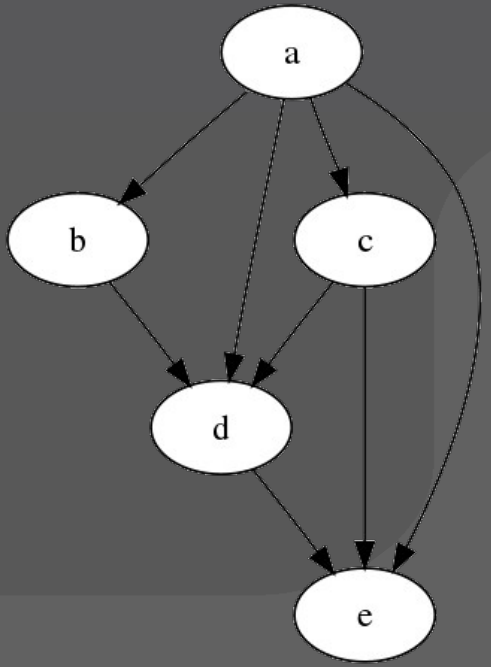
A curated list of awesome open source workflow engines

Full fledged product

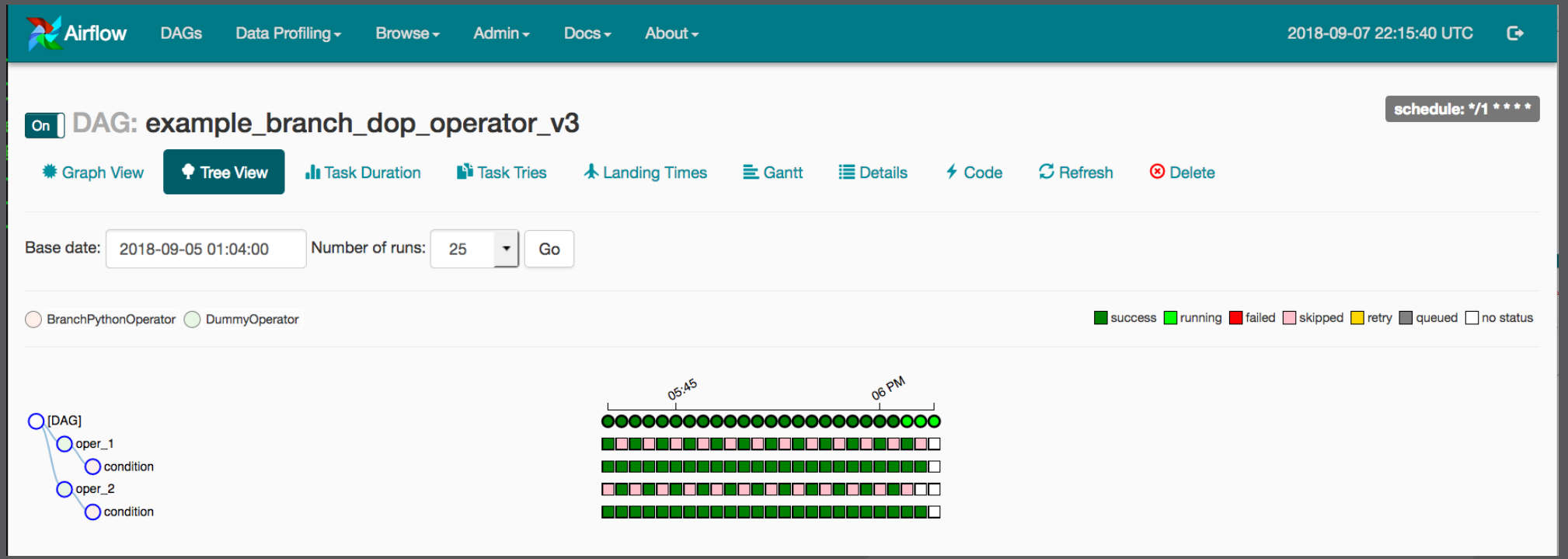
- Airflow stars 27k - Python-based platform for running directed acyclic graphs (DAGs) of tasks
- Argo Workflows stars 12k - Open source container-native workflow engine for getting work done on Kubernetes
- Azkaban stars 4.1k - Batch workflow job scheduler created at LinkedIn to run Hadoop jobs.
- Brigade 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

Kategorie Python / DAG Workflows

- Tasks in Python programmiert
- Workflows als DAG (Directed Acyclic Graph)

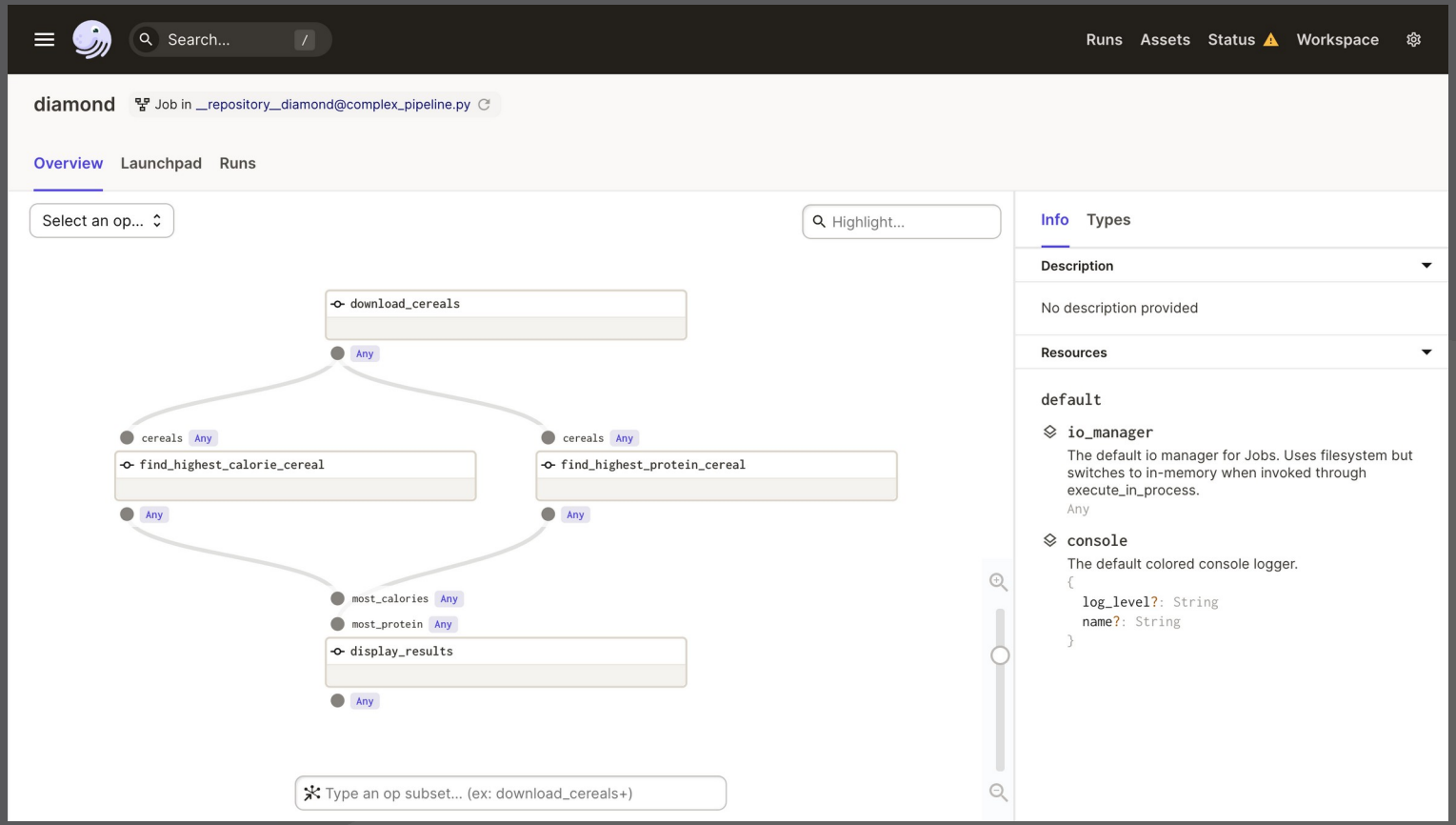


Apache Airflow – DAG (directed acyclic graph)



airflow.apache.org

Dagster

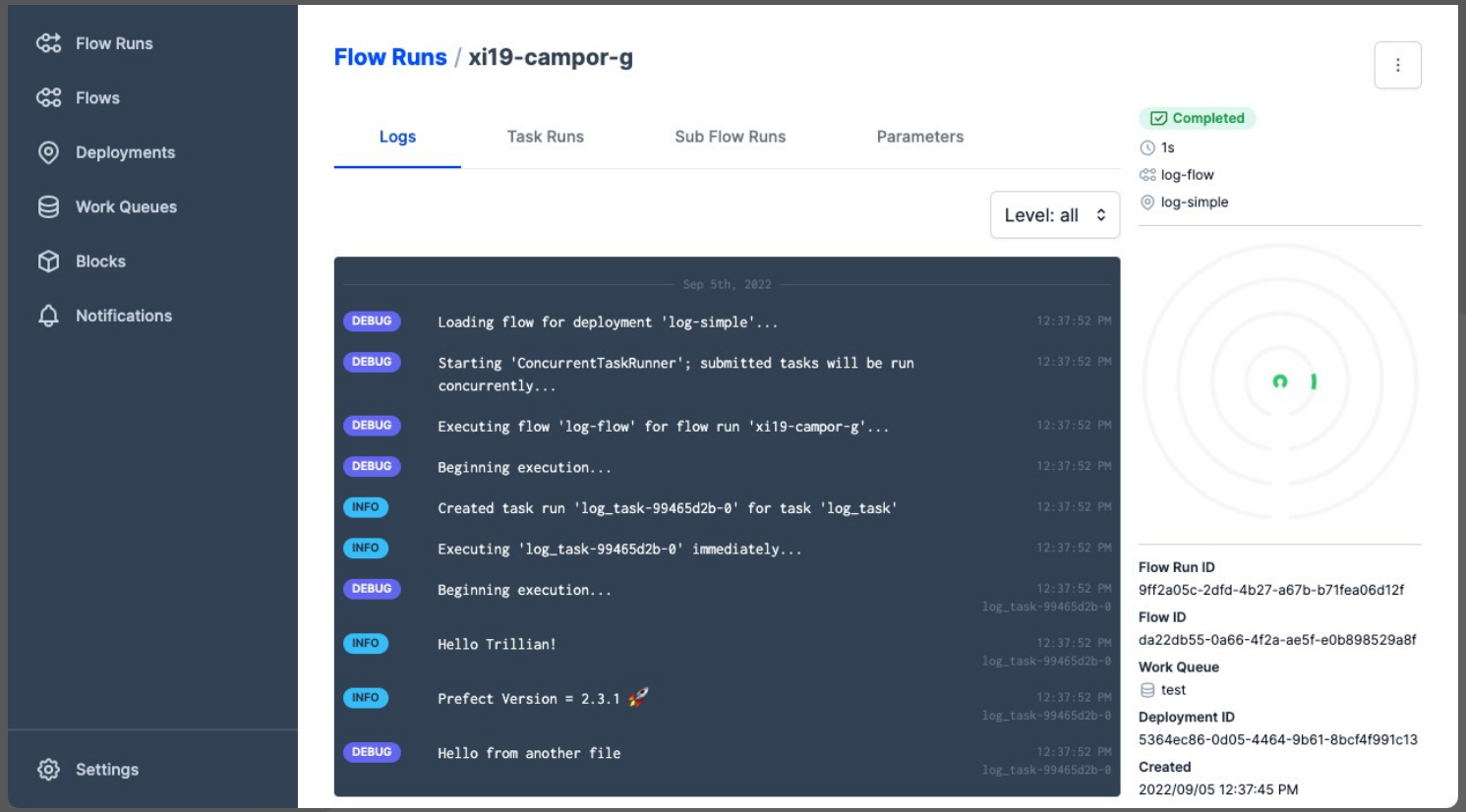


dagster.io

Dagster: Multi backend processing engine

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

Prefect



www.prefect.io

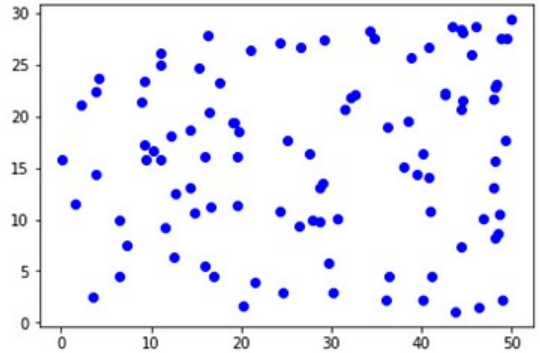
Kategorie Notebooks

- Kombination von Text und Code
- Interaktive Erstellung im Browser

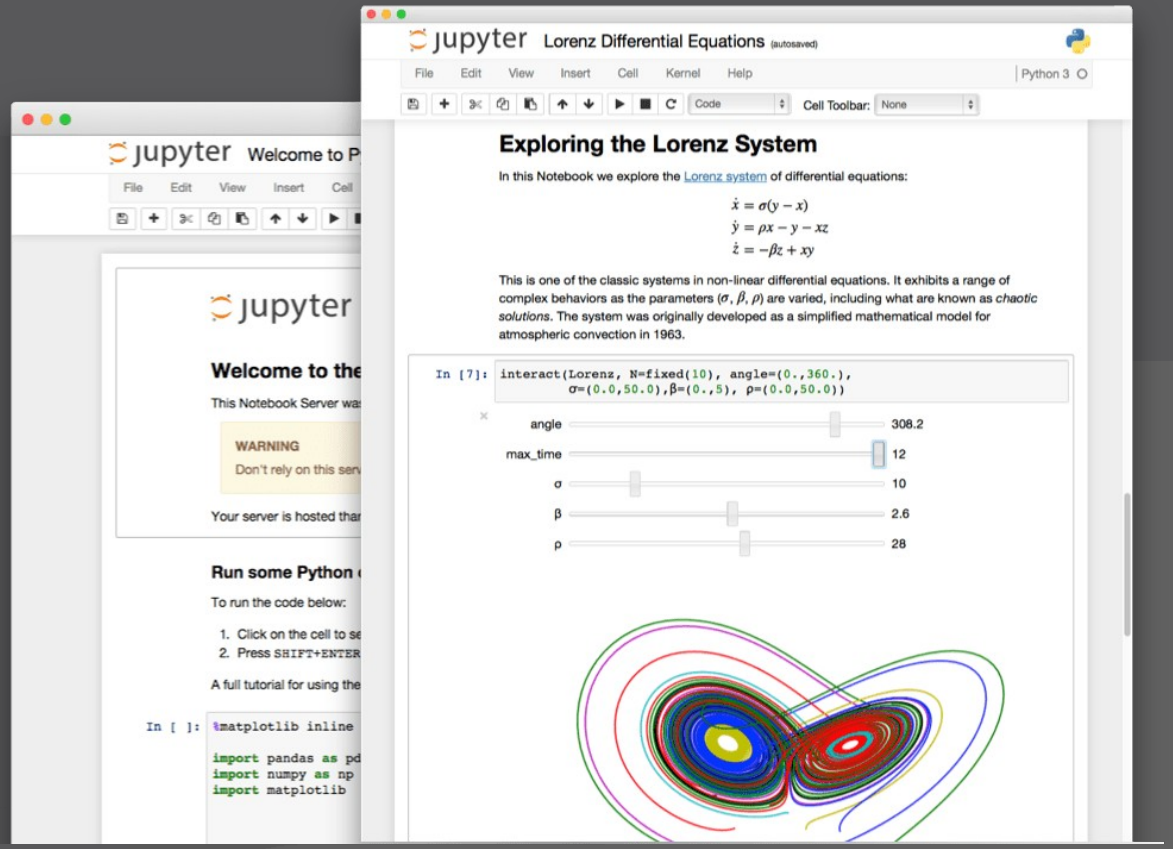
Notes for presentation

In [6]:
```
import numpy as np
import matplotlib.pyplot as plt
X = np.random.uniform(0,50,100)
Y = np.random.uniform(0,30,100)

plt.plot(X,Y,'bo')
plt.show()
```



Jupyter Notebook



jupyter.org

Kategorie SQL Workflows

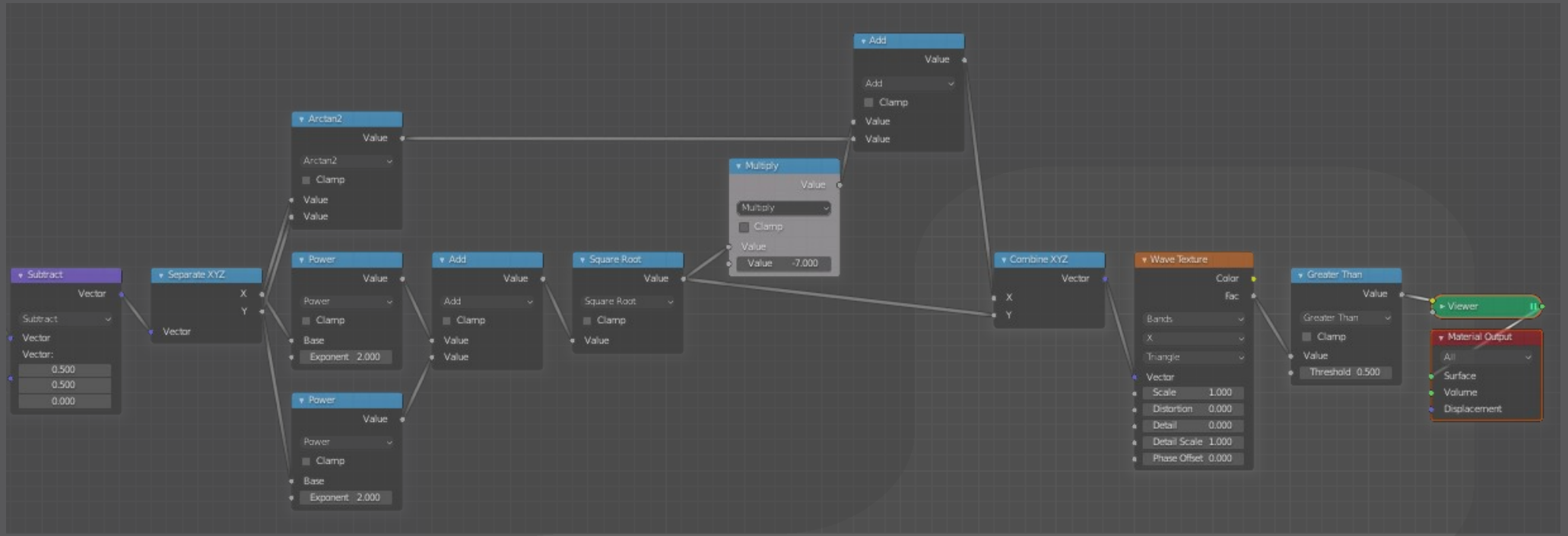
- Datenverarbeitung mit SQL
- Workflows deklarativ

```
INSERT INTO fpdslv95.lfp2_gemeinde(bfsnr, aname, lfp2_von)
SELECT gem.bfs_nummer, gem.name, f.t_id
FROM fpdslv95.lfp2 f,
     fpds2.gemeindegrenzen gem
WHERE ST_Intersects(ST_Buffer(f.geometrie, 0.05), gem.geom);
```

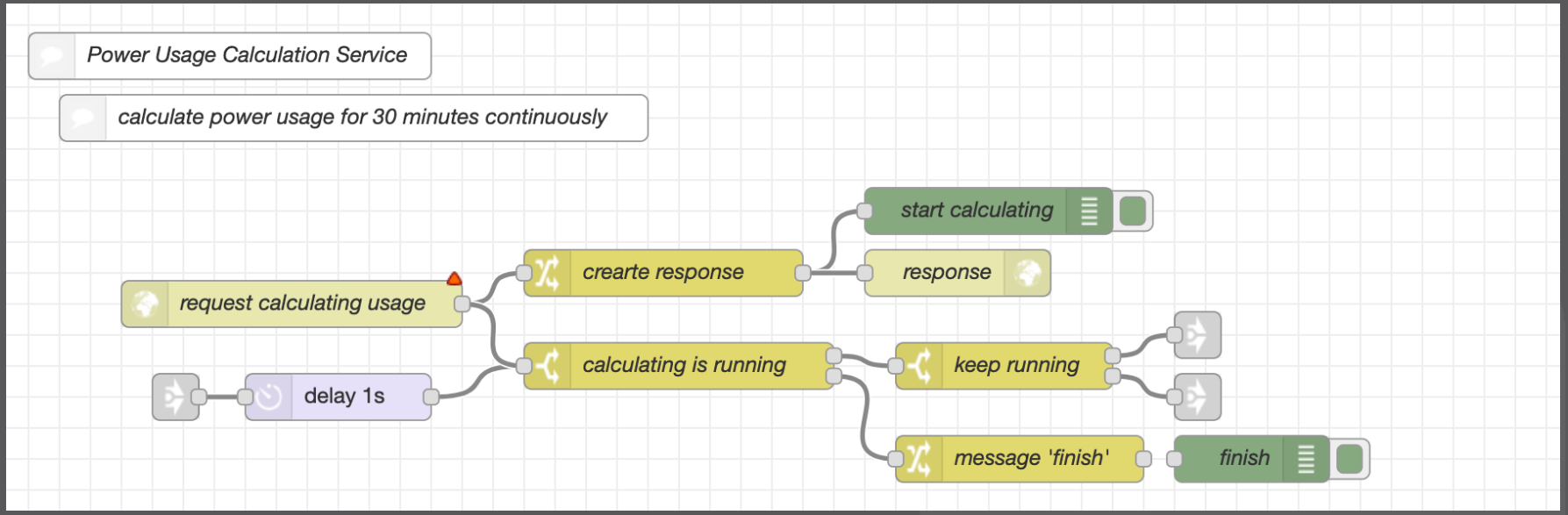

```
jaffle_shop
├── README.md
├── analysis
├── data
│   └── employees.csv
├── dbt_project.yml
├── macros
│   └── cents_to_dollars.sql
├── models
│   ├── intermediate
│   │   └── finance
│   │       ├── _int_finance__models.yml
│   │       └── int_payments_pivoted_to_orders.sql
│   ├── marts
│   │   ├── finance
│   │   │   ├── _finance__models.yml
│   │   │   ├── orders.sql
│   │   │   └── payments.sql
│   │   └── marketing
│   │       ├── _marketing__models.yml
│   │       └── customers.sql
```

www.getdbt.com

Kategorie Low-Code Workflows

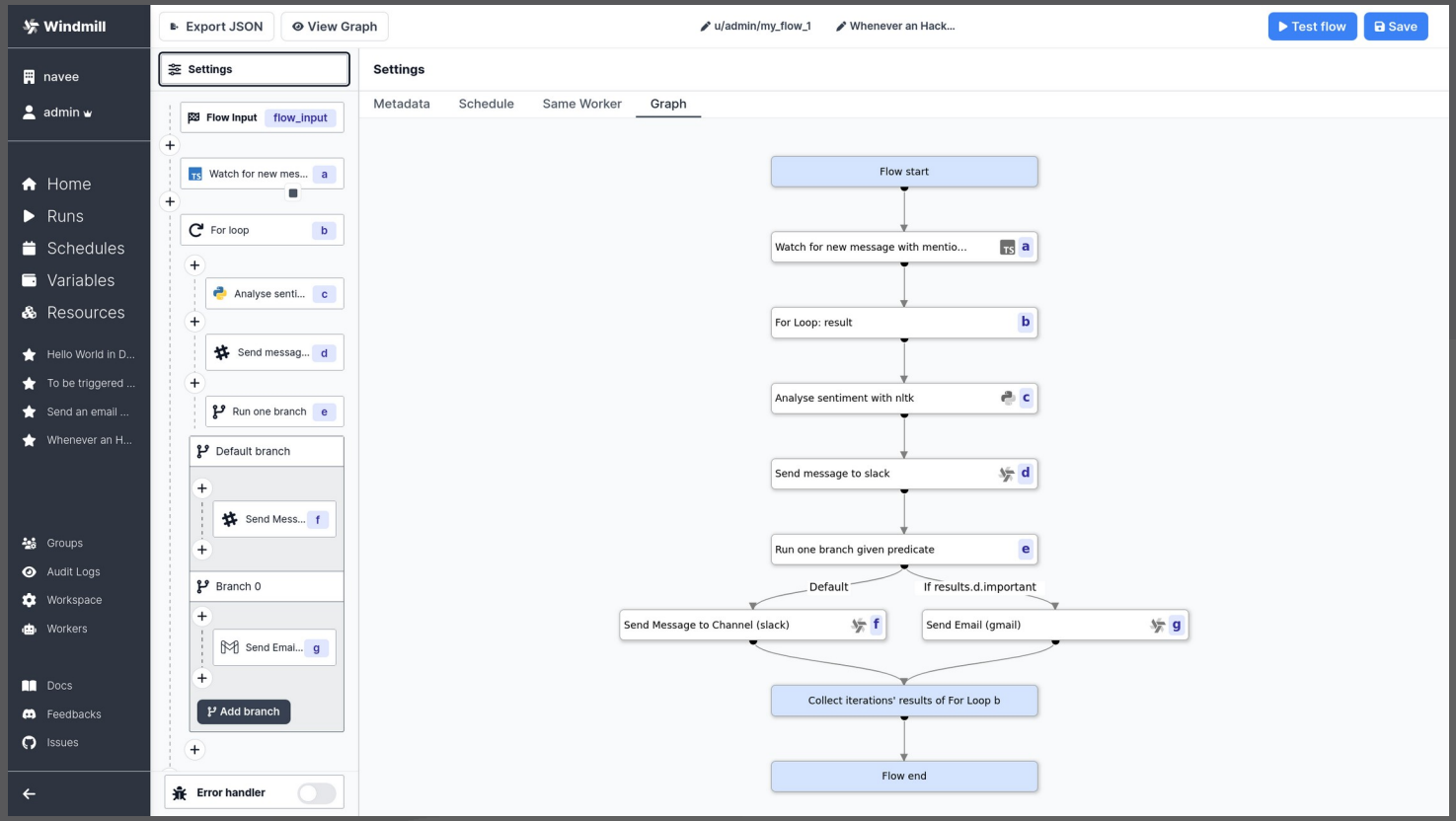


Node-RED



nodered.org

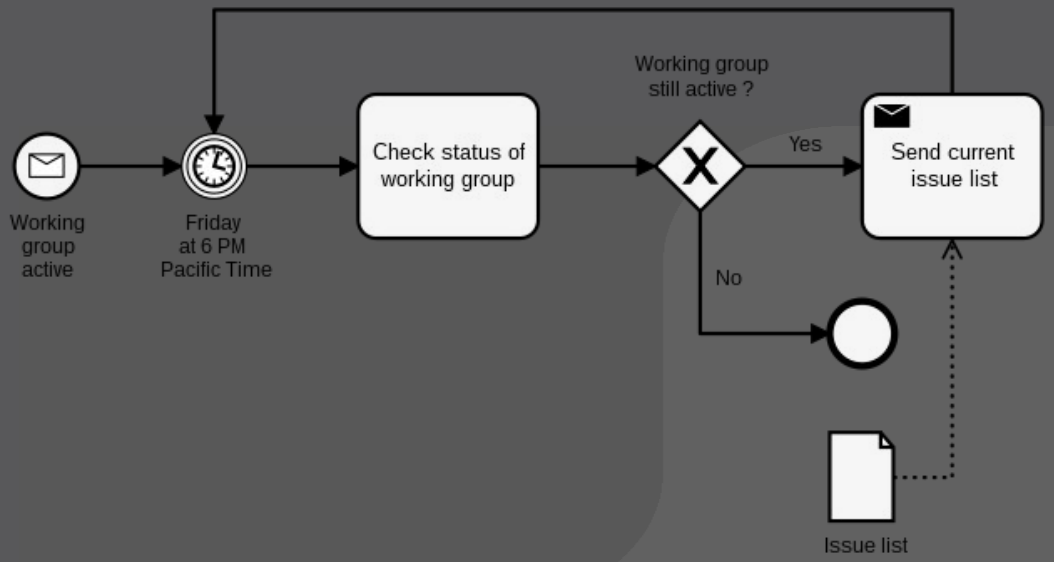
windmill



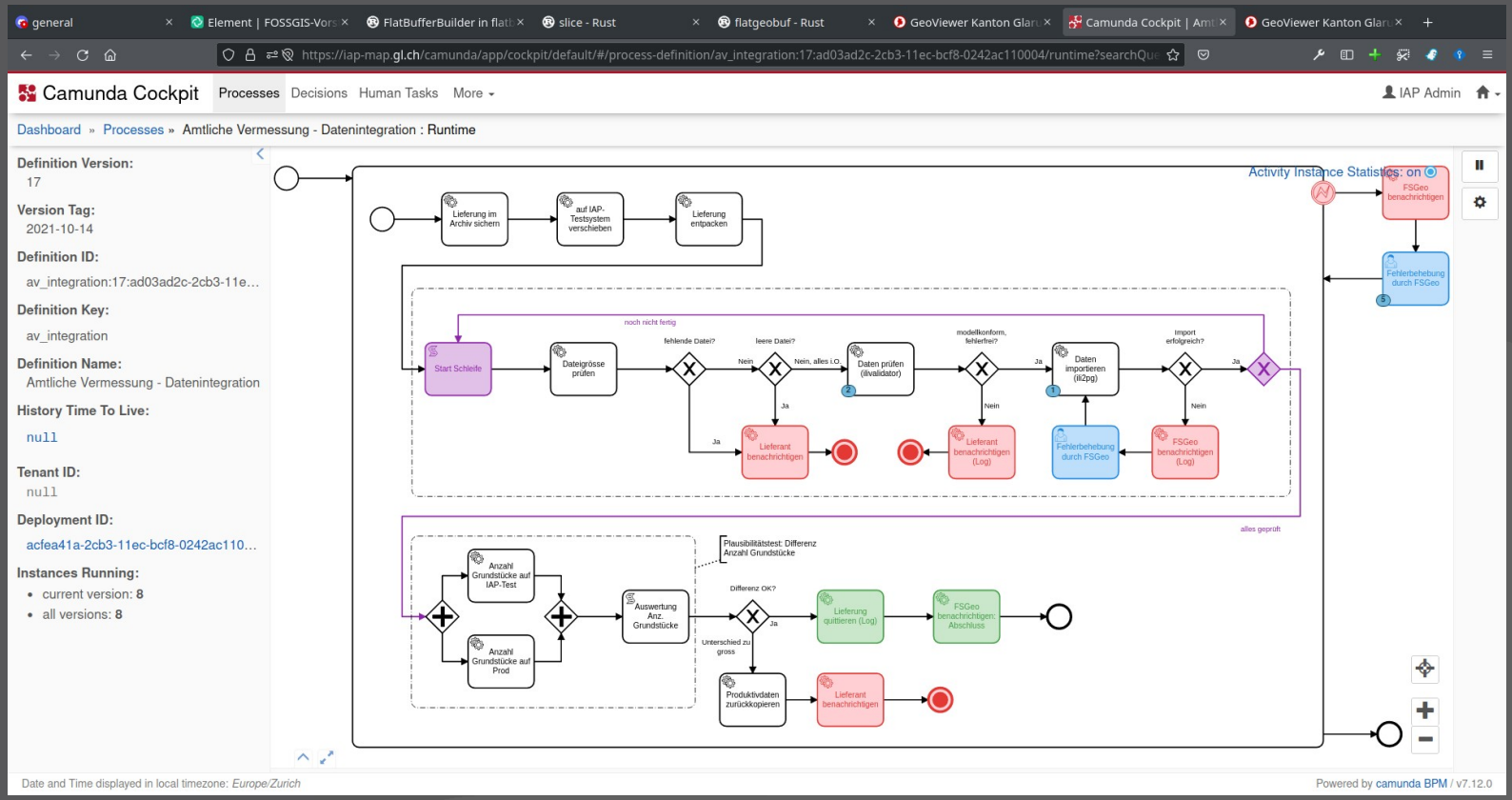
www.windmill.dev

Kategorie Grafische Prozessmodellierung (BPMN)

- BPMN: Business Process Model and Notation



Camunda



camunda.com

Überblick

- **Python / DAG**
  - Apache Airflow
  - Dagster
  - Prefect
- **Notebooks**
  - Jupyter Notebooks
- **SQL**
  - Dbt
- **Low-Code**
  - Node-RED
  - Windmill
- **Grafische Prozess-modellierung (BPMN)**
  - Camunda

OGC API - Processes

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

OGC API - PROCESSES

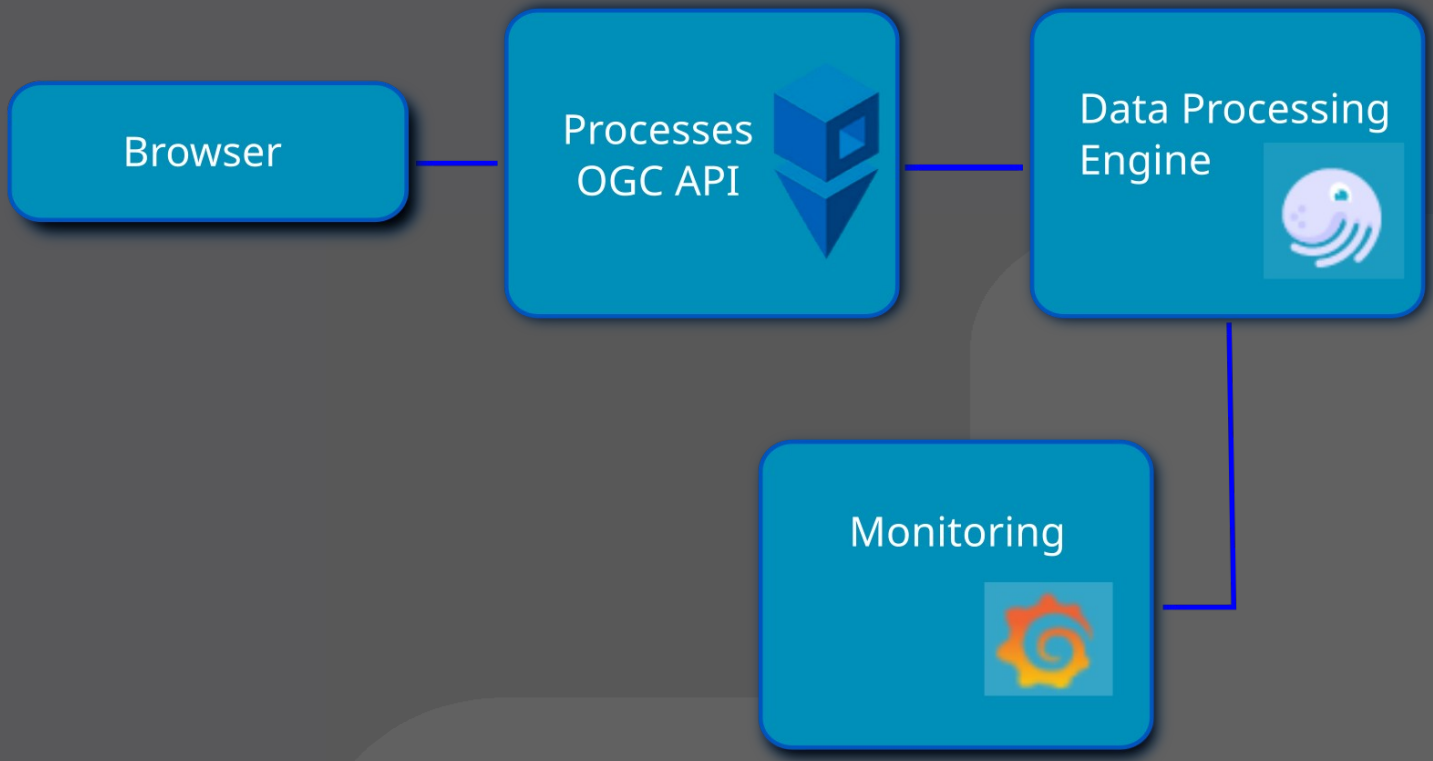
The OGC API - Processes standard supports the wrapping of computational tasks into executable processes that can be offered by a server through a Web API and be invoked by a client application. The standard specifies a processing interface to communicate over a RESTful protocol using JavaScript Object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

Overview

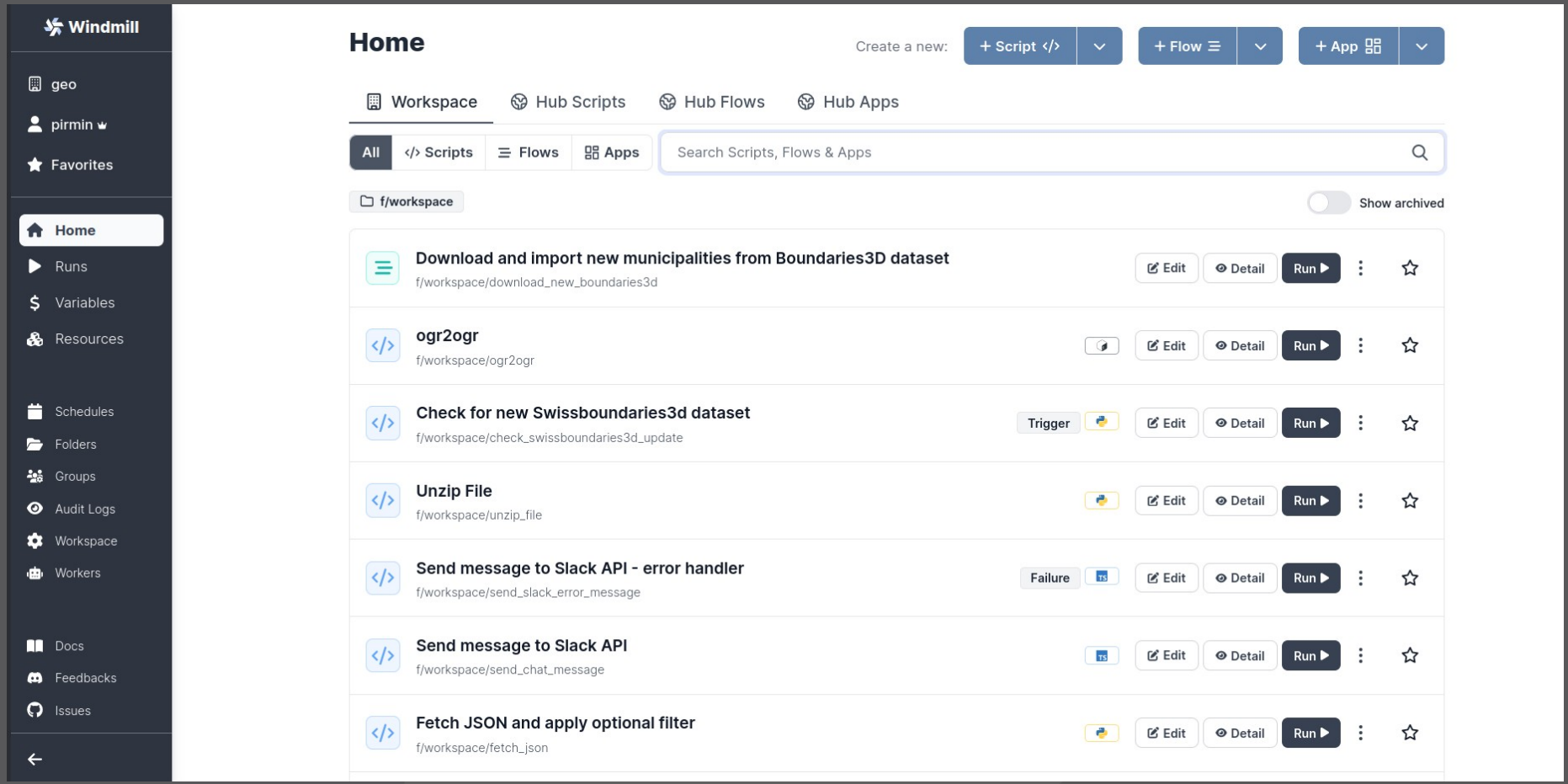
- Visit the overview page to learn more about this API.

Developer-friendly OpenAPI definitions

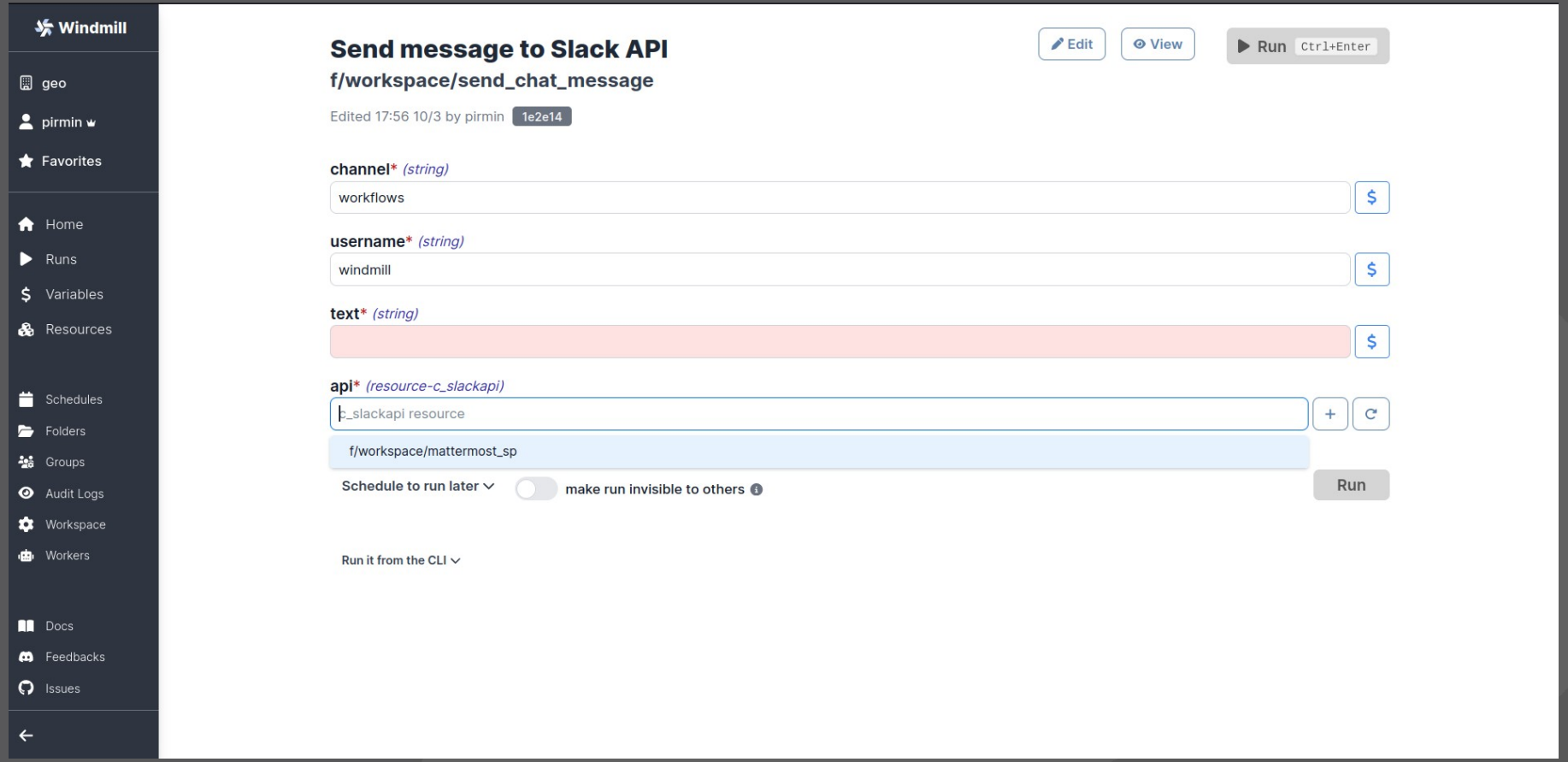
Processing engine + OGC API + Monitoring



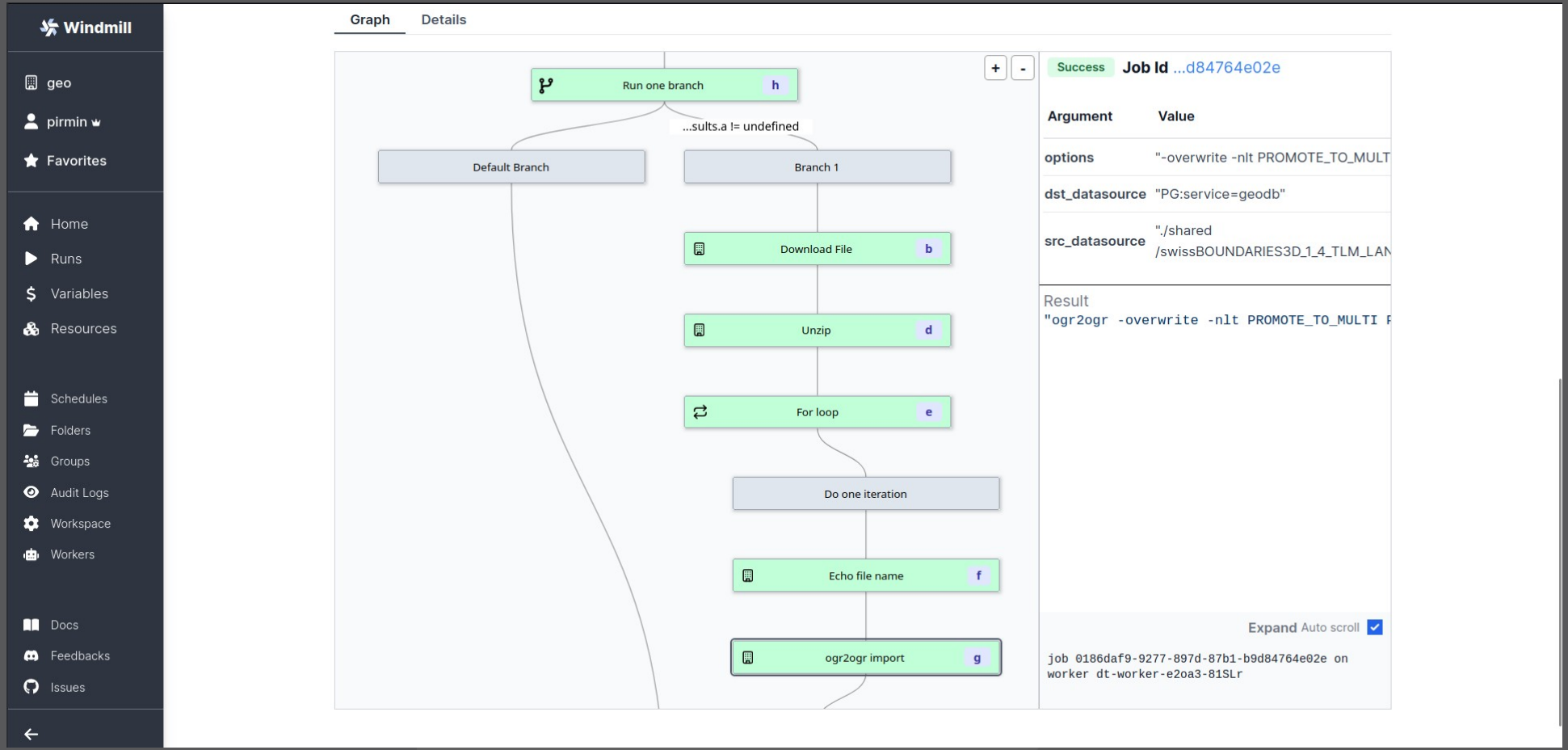
Demo Low-Code Windmill



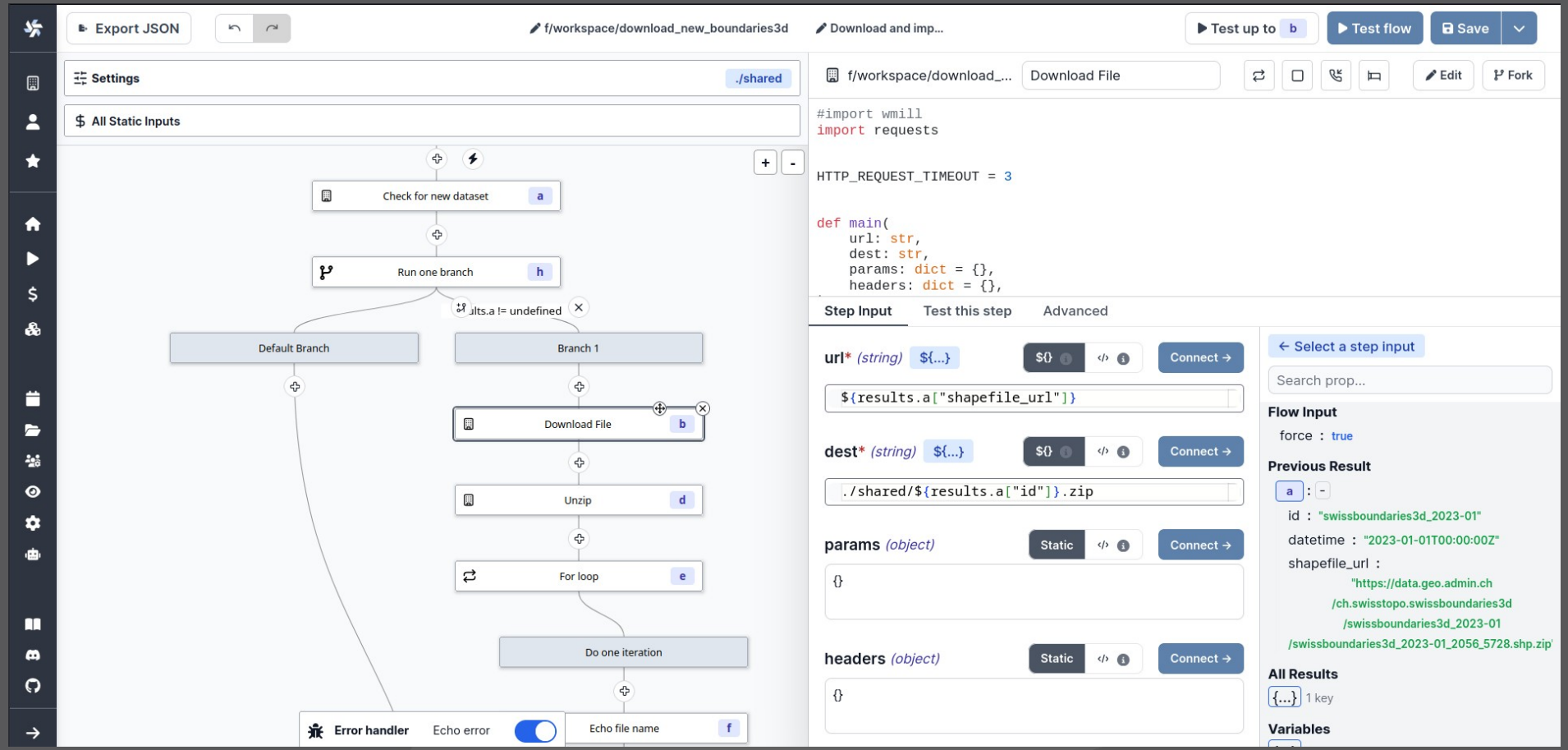
Demo Low-Code Windmill



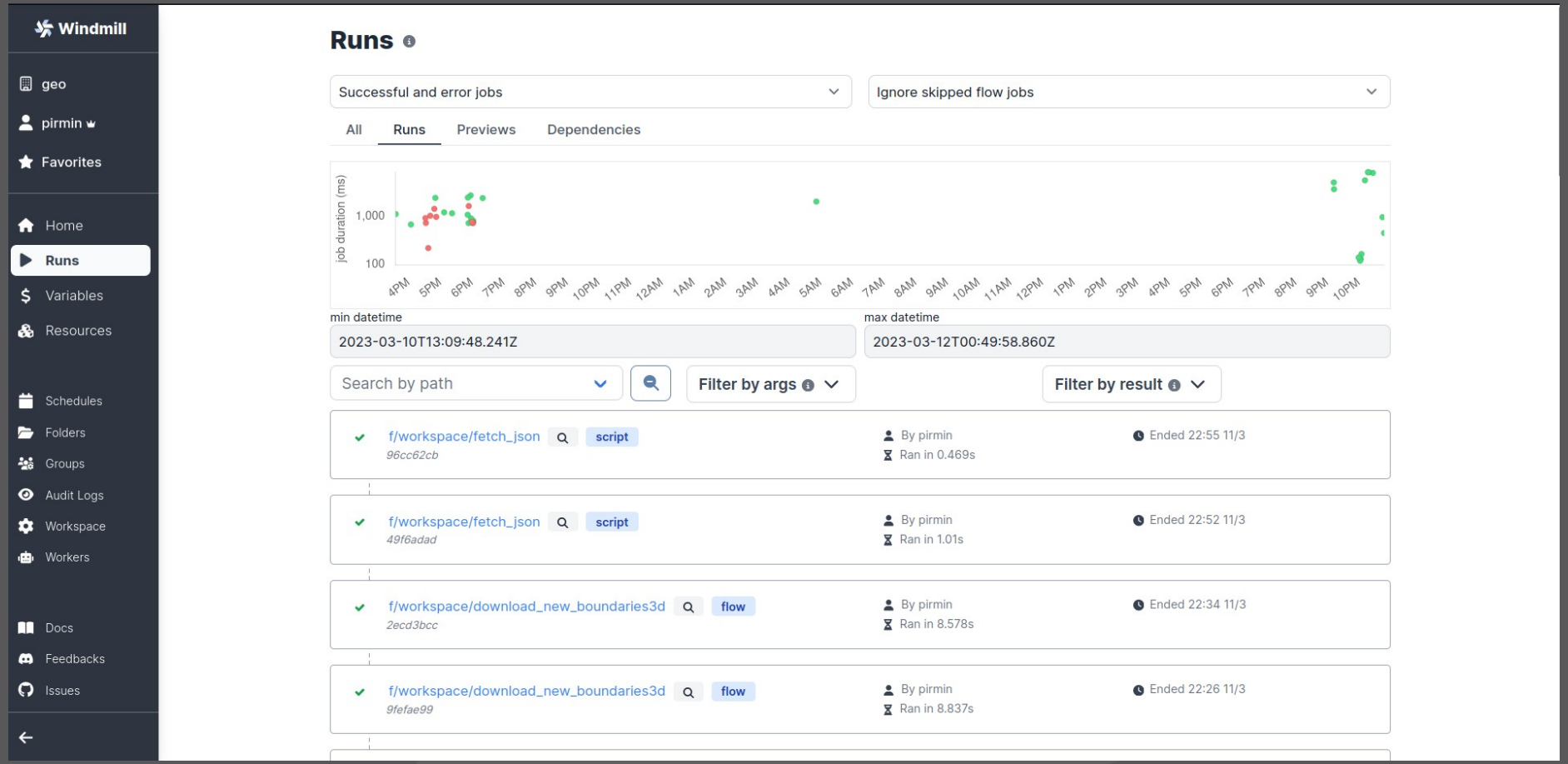
Demo Low-Code Windmill



Demo Low-Code Windmill



Demo Low-Code Windmill



Danke!

Pirmin Kalberer
@implgeo@mapstodon.space