

Streamline data science workflow development using Jupyter Notebook And Airflow

Neha Singla, Apple
Sathish Kumar Thangaraj, Apple

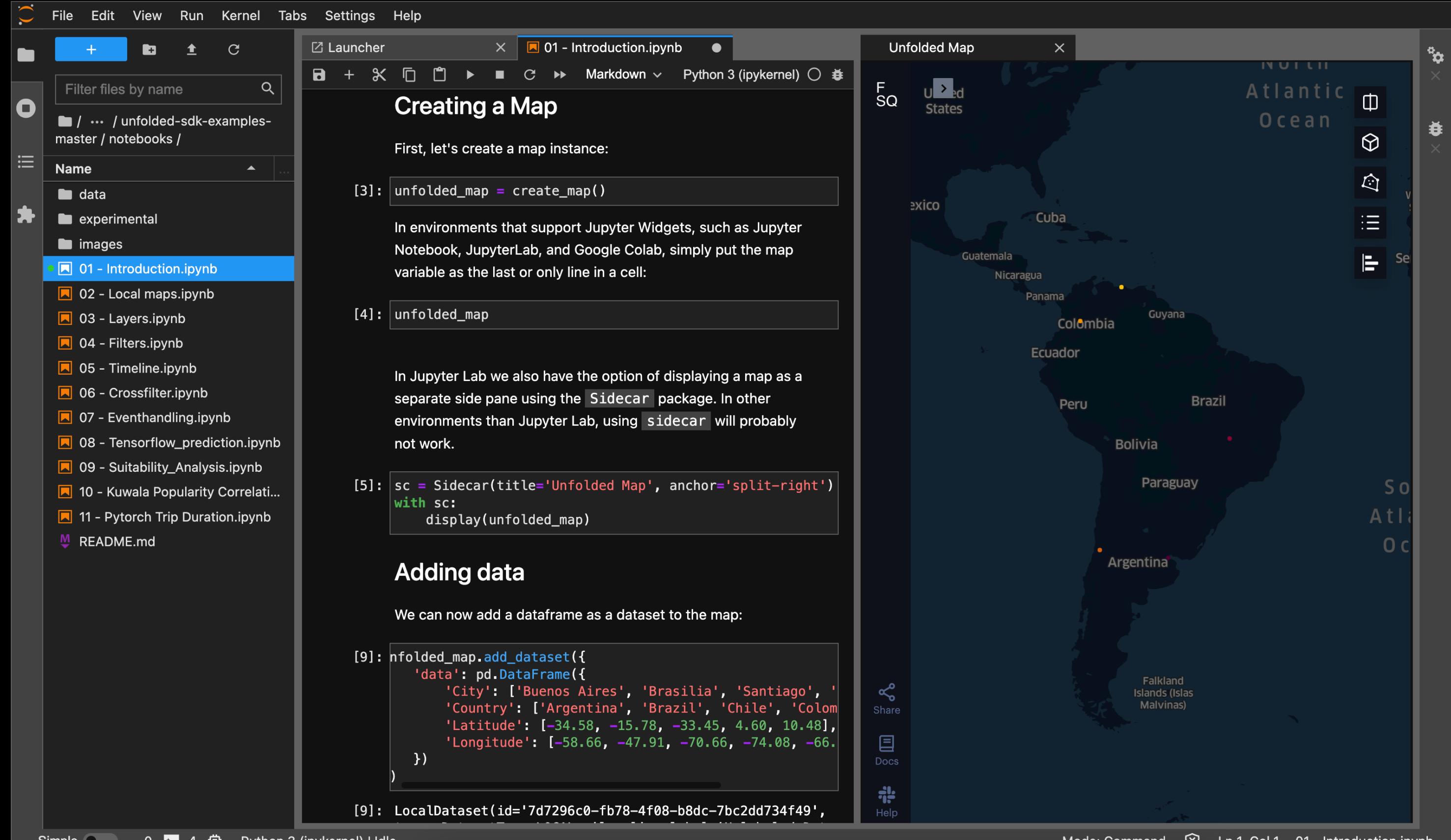
NOT A CONTRIBUTION

Agenda

- Jupyter Notebooks
- Data Science Workflow
- Data Science Workflow- Pain Points
- Streamlining Data Science Workflow
- Demo
- Jupyter Workflow with Airflow Architecture

Jupyter Notebooks

- Jupyter Ecosystem
- Multiple languages
- Prototyping
- Data Exploration
- Iterative Experiments



The screenshot shows the Jupyter Lab interface with a dark theme. On the left, a file browser displays a directory structure with notebooks like '01 - Introduction.ipynb' and other files. The main area shows a notebook cell titled 'Creating a Map' with the following code:

```
[3]: unfolded_map = create_map()  
In environments that support Jupyter Widgets, such as Jupyter Notebook, JupyterLab, and Google Colab, simply put the map variable as the last or only line in a cell:  
[4]: unfolded_map  
In Jupyter Lab we also have the option of displaying a map as a separate side pane using the Sidecar package. In other environments than Jupyter Lab, using sidecar will probably not work.  
[5]: sc = Sidecar(title='Unfolded Map', anchor='split-right')  
with sc:  
    display(unfolded_map)
```

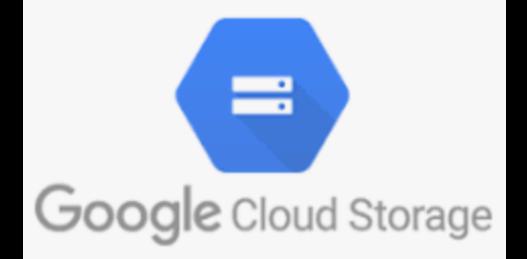
Below this, another cell is shown with the title 'Adding data' and the following code:

```
[9]: unfolded_map.add_dataset({  
    'data': pd.DataFrame({  
        'City': ['Buenos Aires', 'Brasilia', 'Santiago', 'Porto Alegre', 'Rio de Janeiro', 'Sao Paulo'],  
        'Country': ['Argentina', 'Brazil', 'Chile', 'Brazil', 'Brazil', 'Brazil'],  
        'Latitude': [-34.58, -15.78, -33.45, 4.60, 10.48],  
        'Longitude': [-58.66, -47.91, -70.66, -74.08, -66.27]  
    })  
})
```

The right side of the interface features a 'Unfolded Map' pane showing a map of South America and parts of North America and the Atlantic Ocean. The map includes labels for countries and cities.

Data Science Workflow

- Data sources



- Languages, Libraries and frameworks



- Compute

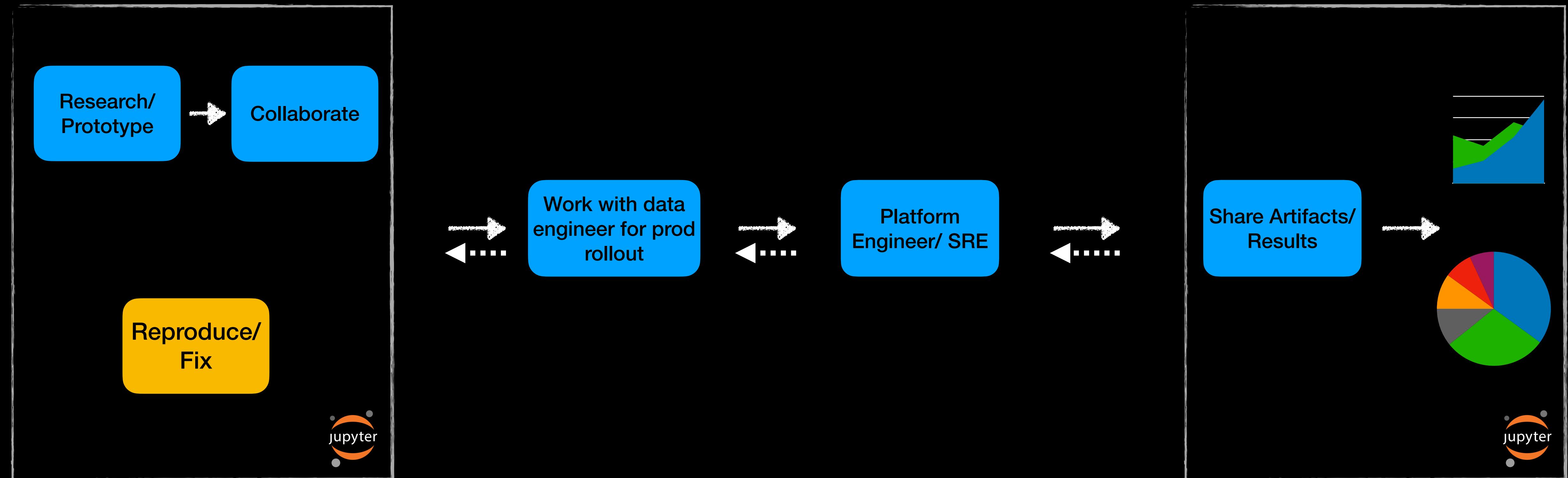


- Machine Learning



- Share and collaborate

Data Science Workflow Continued...



Data Science Workflow- Pain Points

- Knowledge buried in multiple notebook files
- Extract code and work with data engineer to run in production
- Reproducibility
- Running large scale interactive experiments
- Debug failed pipelines

Jupyter - Beyond Interactive use cases

- Long running sessions
- Jobs running at a regular interval
- Notebook workflows

Meet **Workflows** on Jupyter Notebooks.

Workflows > gv2fol4m2vus

Yelp-review-data-ml-workflow

Runs Tasks

```
graph LR; A[Spark-Ingest-Data] --> B[Spark-Feature-Gen]; B --> C[fine-tune-LLM1]; C --> D[fine-tune-LLM]; D --> E[evaluate-LLMs];
```

Workflow details

Name: Yelp-review-data-ml-workflow

Created at: 4/18/2024, 3:08:58 PM

Updated at: 4/18/2024, 3:11:29 PM

Last deployed: 4/18/2024, 3:13:49 PM

Namespace: [REDACTED]

Status: ACTIVE

Deployment status: UPDATED

Demo

Why Airflow ?

- General purpose orchestration system
- Extendibility
- Monitoring capabilities
- Huge community and support

Papermill Operator

- Parameterize Jupyter Notebooks
- Execute Jupyter Notebooks

Papermill Operator Limitations

- Tightly coupled with python runtime
- Scaling
- Multi-tenancy

Papermill Operator- Enhancements

<https://github.com/apache/airflow/pull/34840>

- Support multiple runtimes
- Language Agnostic
- Kernels can run in any cloud

The screenshot shows a GitHub pull request page for the Apache Airflow repository. The pull request is titled "Extend papermill operator to support remote kernels #34840". It has been merged by `bolkedebruin` on Nov 13, 2023. The PR has 40 conversations, 1 commit, 71 checks, and 14 files changed. The description of the PR states:

This PR adds support to run papermill operator that can connect to kernels managed externally by other systems. This would be useful to run the operator in cloud environments and would also be helpful to run spark or scala notebooks.

It extends papermill to support new engine using the entry_points as described [here](#).

It adds unittest and also a system test to run in CI environments.

Validated using below steps in breeze environment:

```
* breeze ci-image build --upgrade-to-newer-dependencies
* breeze start-airflow
* pytest --system papermill tests/system/providers/papermill/example_papermill_remote_verify.py
```

Contributor `akshaychitneni` commented on Oct 9, 2023:

Congratulations on your first Pull Request and welcome to the Apache Airflow community! If you have any issues or are unsure about anything please check our Contribution Guide (<https://github.com/apache/airflow/blob/main/CONTRIBUTING.rst>)

Here are some useful points:

Reviewers: `bolkedebruin`, `Taragolis`

Assignees: None assigned

Labels: `area:providers`, `area:system-tests`, `changelog:skip`, `provider:papermill`

Projects: None yet

Milestone: Airflow 2.8.0

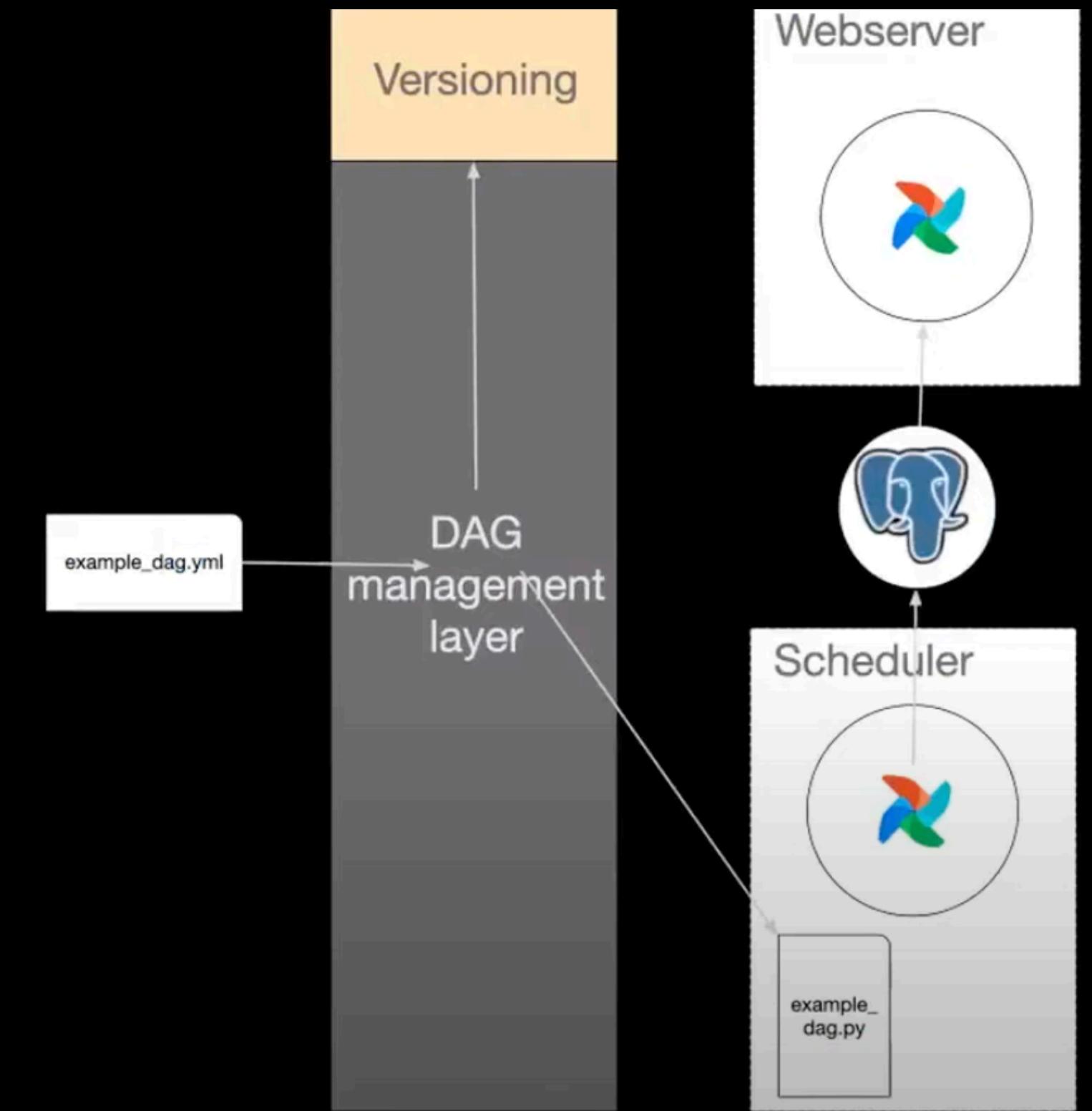
Development: Successfully merging this pull request may close these issues.

None yet

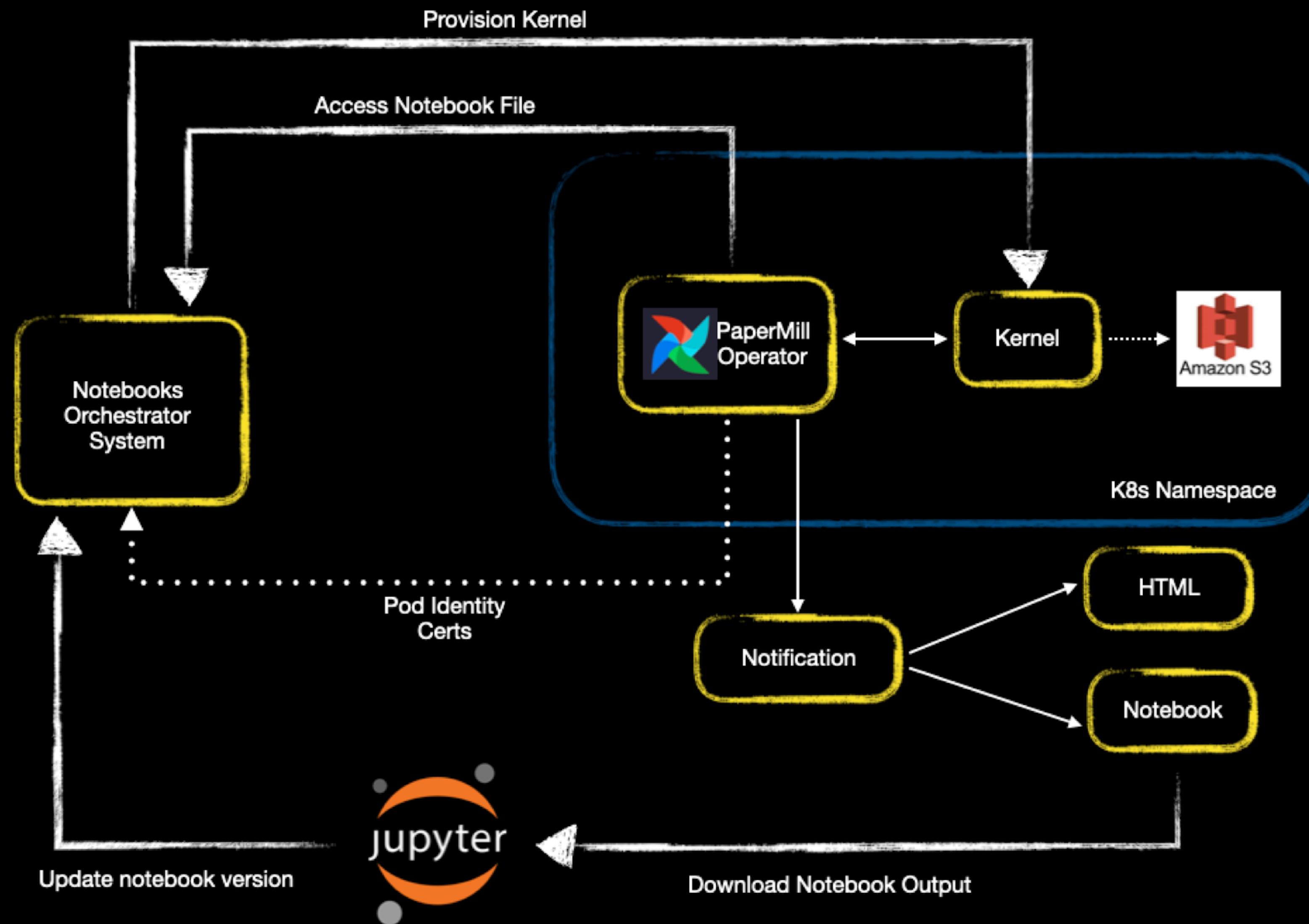
Notifications: [Subscribe](#)

Multi-tenant Pipeline Orchestration using Airflow

- Pipeline Domain Specific Language (DSL)
- Directed Acyclic Graph (DAG) Management
 - Version control
 - Deployment to K8s Namespace/Cluster
- Multi-tenancy
- Airflow Operator marketplace

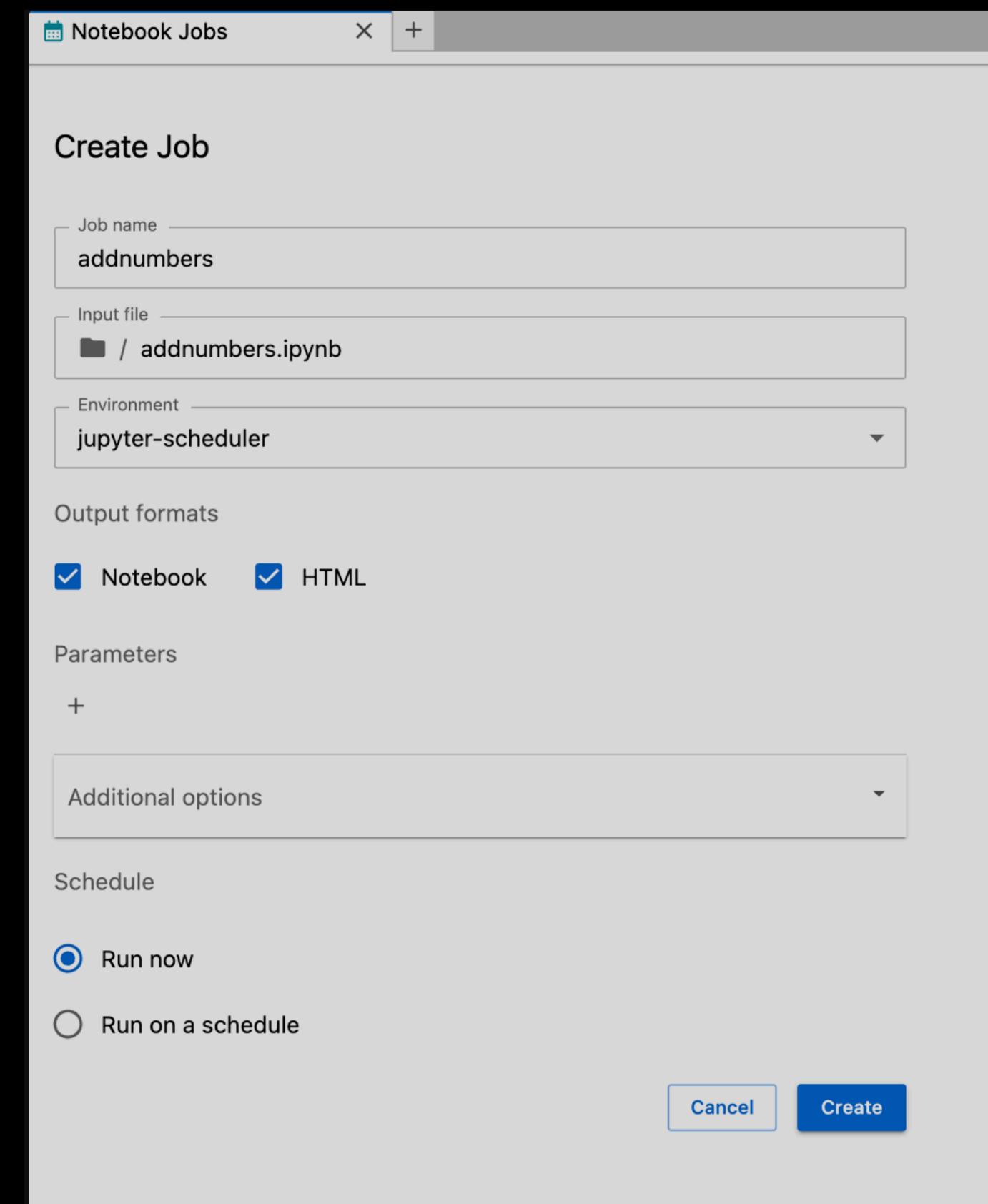
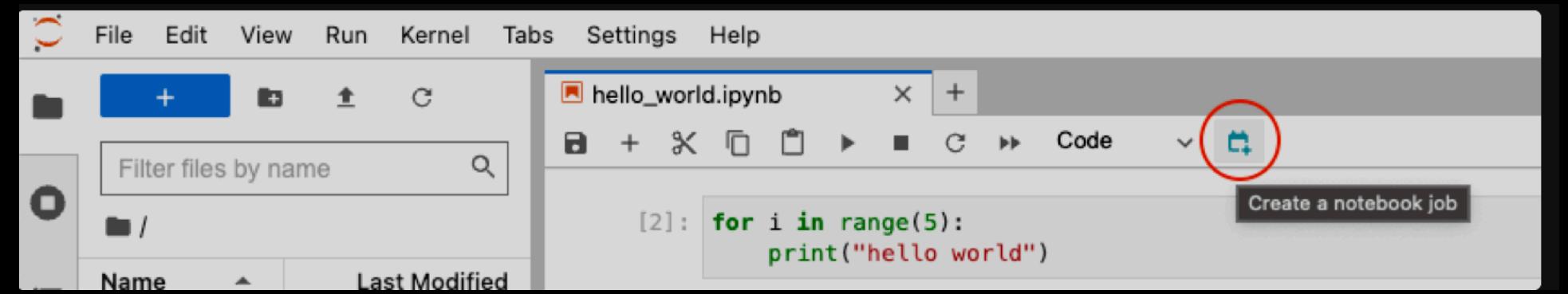


Notebook Execution Using Airflow



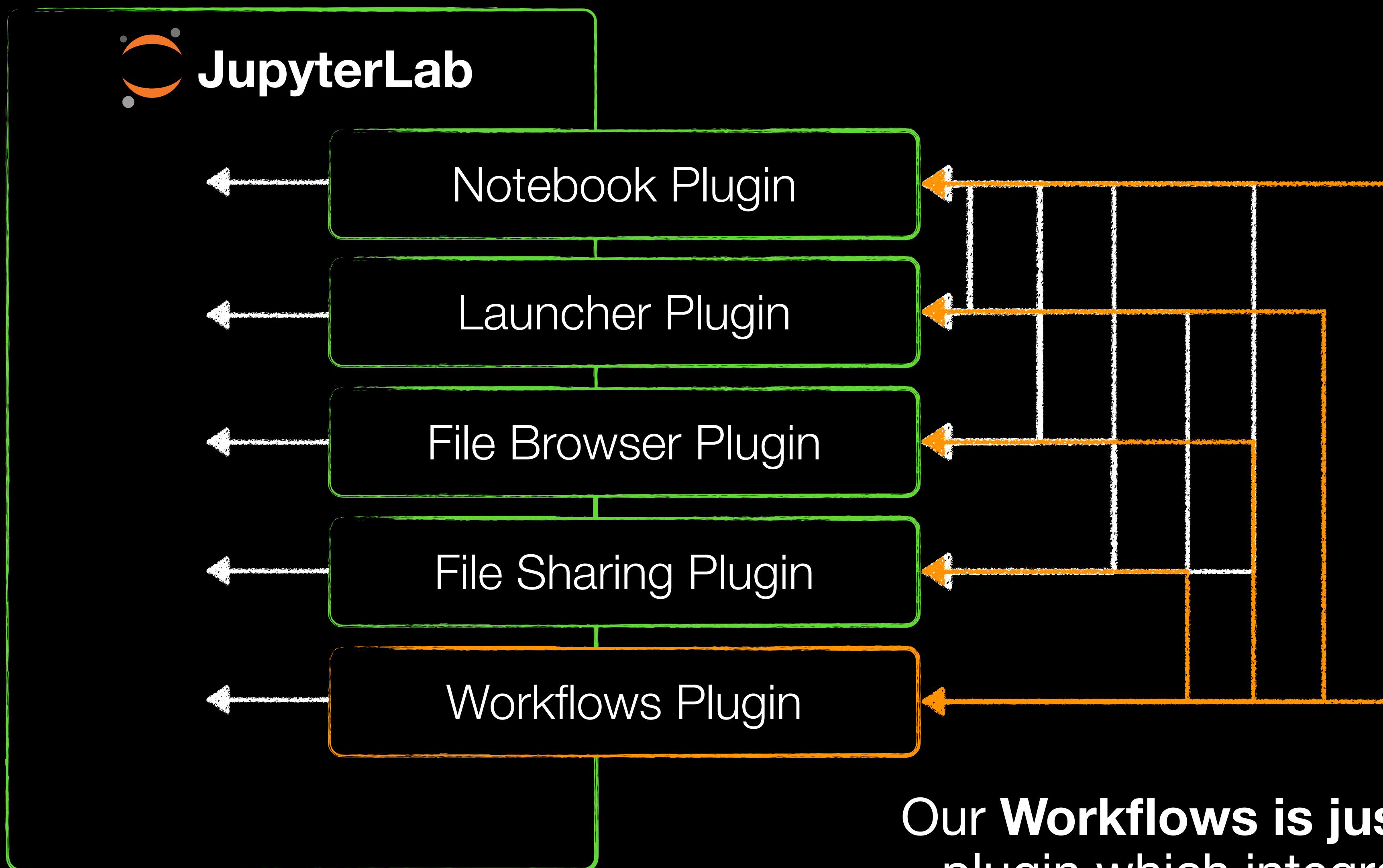
Jupyter Scheduler

- Run jobs in background from Jupyter workspace
 - Run Once
 - Run on a Schedule
- Manage job definitions and jobs
- View historical runs
- Download job run output in multiple formats
- Notebooks as template



Notebook Jobs					
Job name	Input file	Output files	Created at	Status	Actions
sleep	sleep.ipynb	Notebook HTML	12/1/2022, 5:02:25 PM	Completed	X

Dynamic DAG Creation in Jupyter



Our **Workflows** is just another plugin which integrates with many other core plugins.

Future Work

- Open Source
 - Inviting Collaborators
 - Jupyter Meetings
 - <https://jupyter-server.readthedocs.io/en/latest/contributors/team-meetings.html>
- Features
 - Event Driven Notebook workflows
 - Workflow Sharing

Thank You

Questions?