Setting Data & ML Pipeline with GitHub Actions

Data Innovation Summit 2025

Rami Krispin, May 7th, 2025

Agenda

- Motivation
- What GitHub Actions Is
- Data & ML Pipelines with GitHub Actions

Rami Krispin

Senior Manager - Data Science & Engineering

Author | Open Source | Docker Captain 🐳

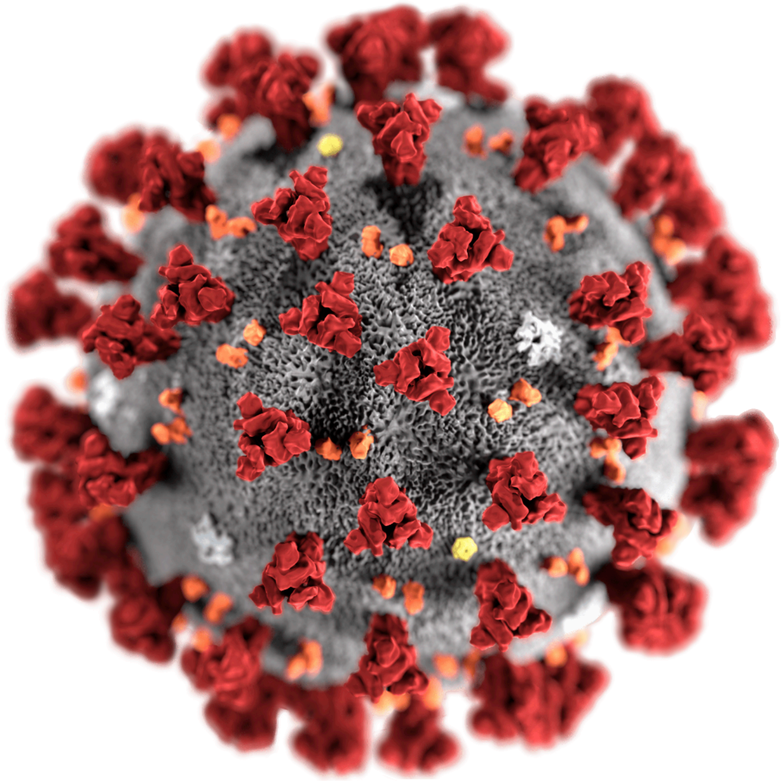
Motivation

coronavirus

The coronavirus package provides a tidy format for the COVID-19 dataset collected by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The dataset includes daily new and death cases between January 2020 and March 2023 and recovery cases until August 2022.

More details available here, and a `csv` format of the package dataset available here

Data source: https://github.com/CSSEGISandData/COVID-19



Source: Centers for Disease Control and Prevention's Public Health Image Library

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Citation

Citing coronavirus

Developers

Rami Krispin
Author, maintainer

Jarrett Byrnes
Author

Dev status

R-CMD

Data Pipeline passing

CRAN 0.4.1

lifecycle stable

License MIT

last commit march

downloads 74K

Connecting through Power BI 1
#14 by jimmyd7377 was closed on Mar 20, 2020

Error when devtools::install_github("RamiKrispin/coronavirus") 1
#13 by Danielchui was closed on Mar 20, 2020

Error in Hubei data for 3/11/2020 2
#12 by tcarleton was closed on Mar 13, 2020

Country name in standard format 12
#11 by shubhrampandey was closed on Jul 1, 2020

0 Case Vectors (non-essential) 2
#10 by j4yr0u93 was closed on Mar 13, 2020

How do u make it realtime and auto update 3
#9 by navmedvideos was closed on Apr 25, 2020

Negative values were found in the package 7
#7 by ddong63 was closed on Mar 10, 2020

Update package locally 2
#6 by shubhrampandey was closed on Mar 13, 2020

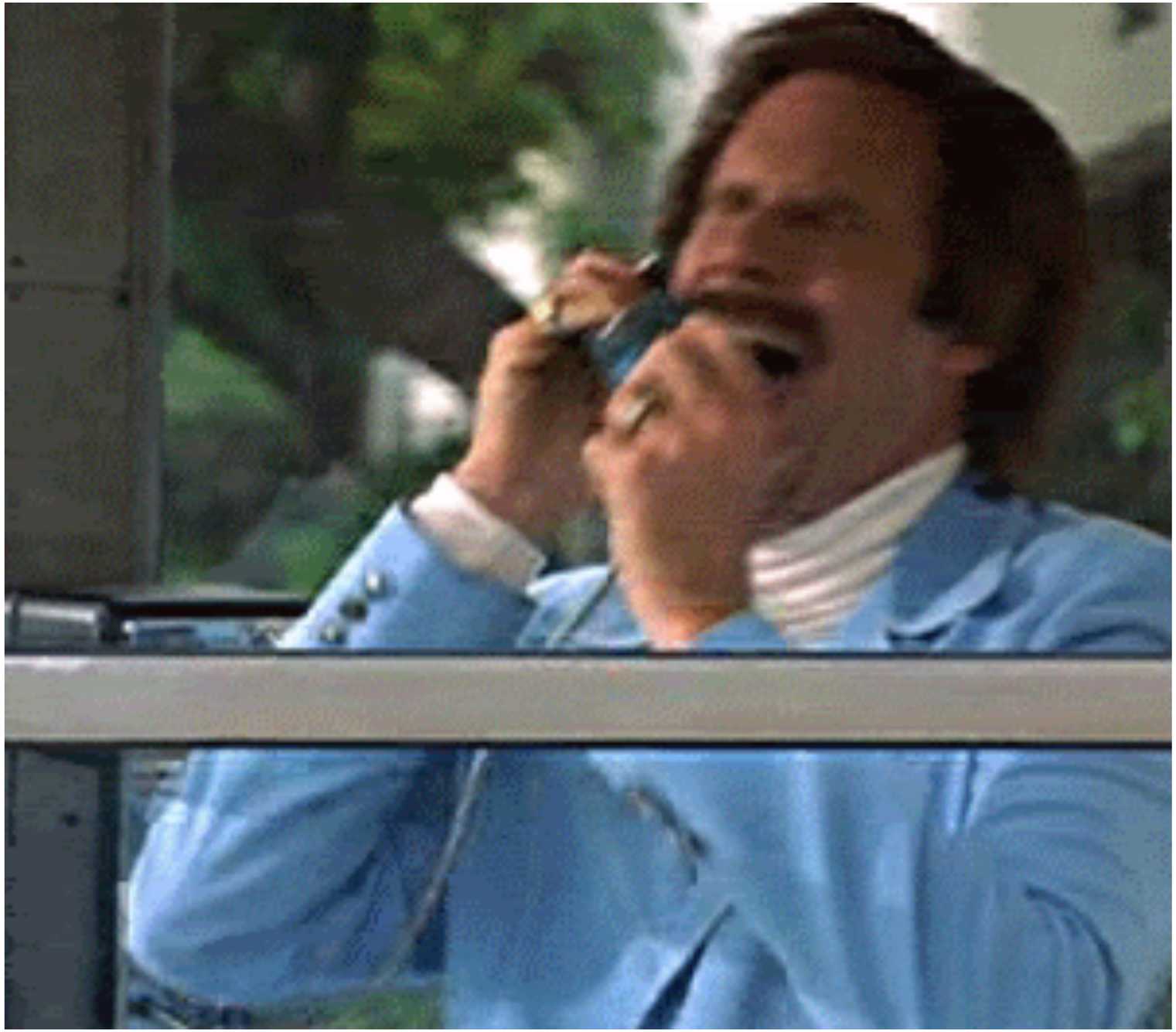
Negative case values 2
#5 by j4yr0u93 was closed on Mar 10, 2020

Covid-19 3
#4 by acgerstein was closed on Mar 6, 2020

Adding a country filter to the dashboard 3
#3 by Agusum was closed on Mar 8, 2020

" Azerbaijan" Has Space in the Name 1
#2 by cannin was closed on Feb 29, 2020

is there any plan to automate data update? 16
#1 by statkclee was closed on Apr 25, 2020



GitHub Actions

A Platform to Automate Workflows

ai-dev-2024-ml-workshop

Public

generated from RamiKrispin/vscode-python-template

Pin | Watch 3 | Fork 18 | Star 93

main | 4 Branches | 0 Tags | Go to file t | Add file | Code

RamiKrispin Auto update of the data ✓ 6578d82 · 3 hours ago | 2,129 Commits

.devcontainer	updated the dev containers settings	11 months ago
.github/workflows	Update data_refresh.yml	10 months ago
.vscode	add docker settings and update the dev contianer settings	11 months ago
data	Auto update of the data	3 hours ago
diagrams	updated the diagram	10 months ago
docker	update the docker settings, add the gt library	11 months ago
docs	Auto update of the data	3 hours ago
functions	fixed the query start and end arguments	11 months ago
images	updated the readme	11 months ago
mlruns/0	add mlflow experiments	11 months ago
prototype	added a function to forecast plot	11 months ago
settings	refreshed the forecast	11 months ago
slides	added the workshop slides	11 months ago
.gitignore	removed the mlruns folder from git	11 months ago
README.md	Update README.md	11 months ago

About

Materials for the AI Dev 2024 conference workshop "Deploy and Monitor ML Pipelines with Python, Open Source, and Free Applications"

ramikrispin.github.io/ai-dev-2024-ml-...

python | docker | workshop | mlops | github-actions

Readme
Activity
93 stars
3 watching
18 forks

Releases

No releases published
Create a new release

Packages

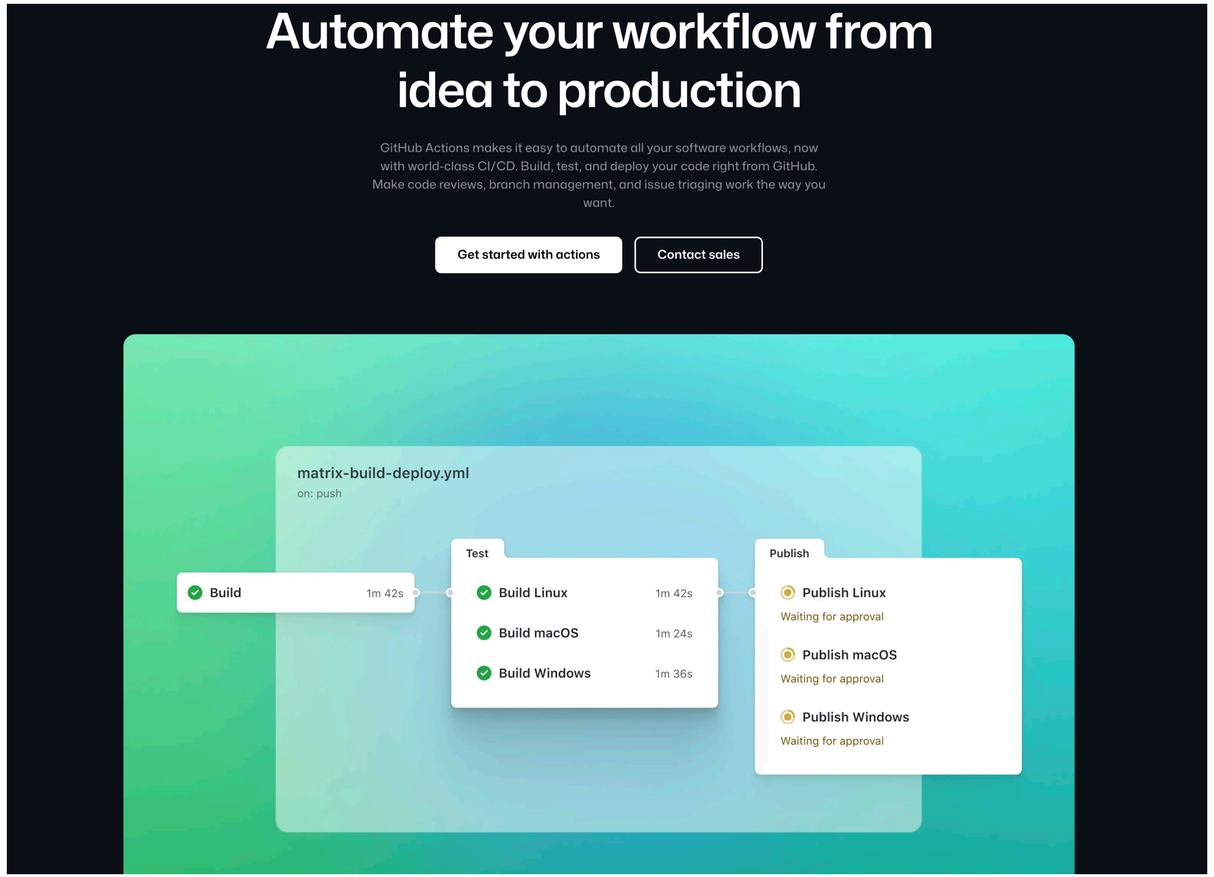
No packages published
Publish your first package

Contributors 2

RamiKrispin Rami Krispin

Workflow Automation

- Unit testing
- Code deployment
- Matrix build
- General purpose automation



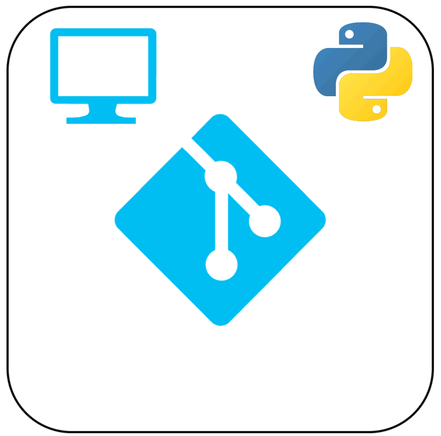
Workflow Types

Triggered

Scheduled

Triggered Workflow

Code Testing

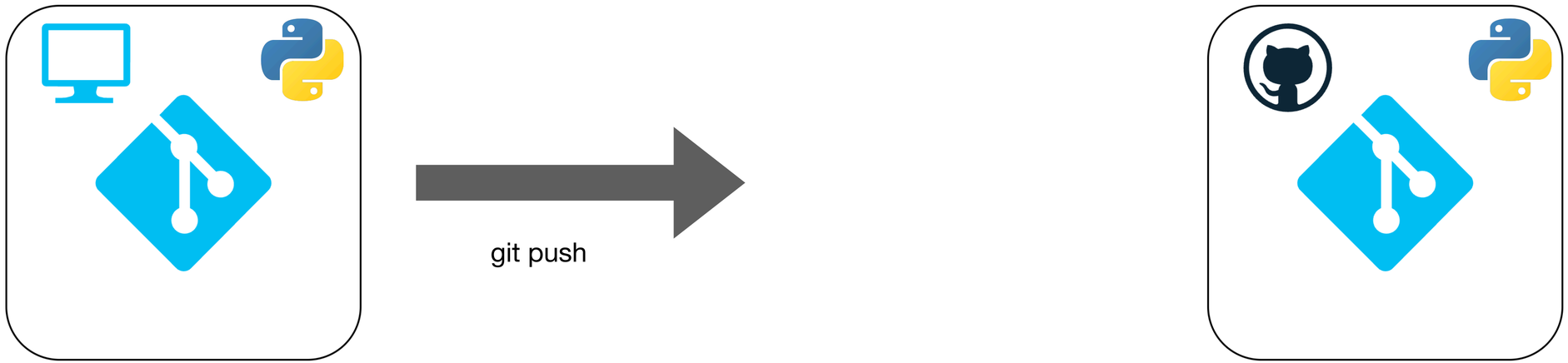


Local

Remote

Triggered Workflow

Code Testing

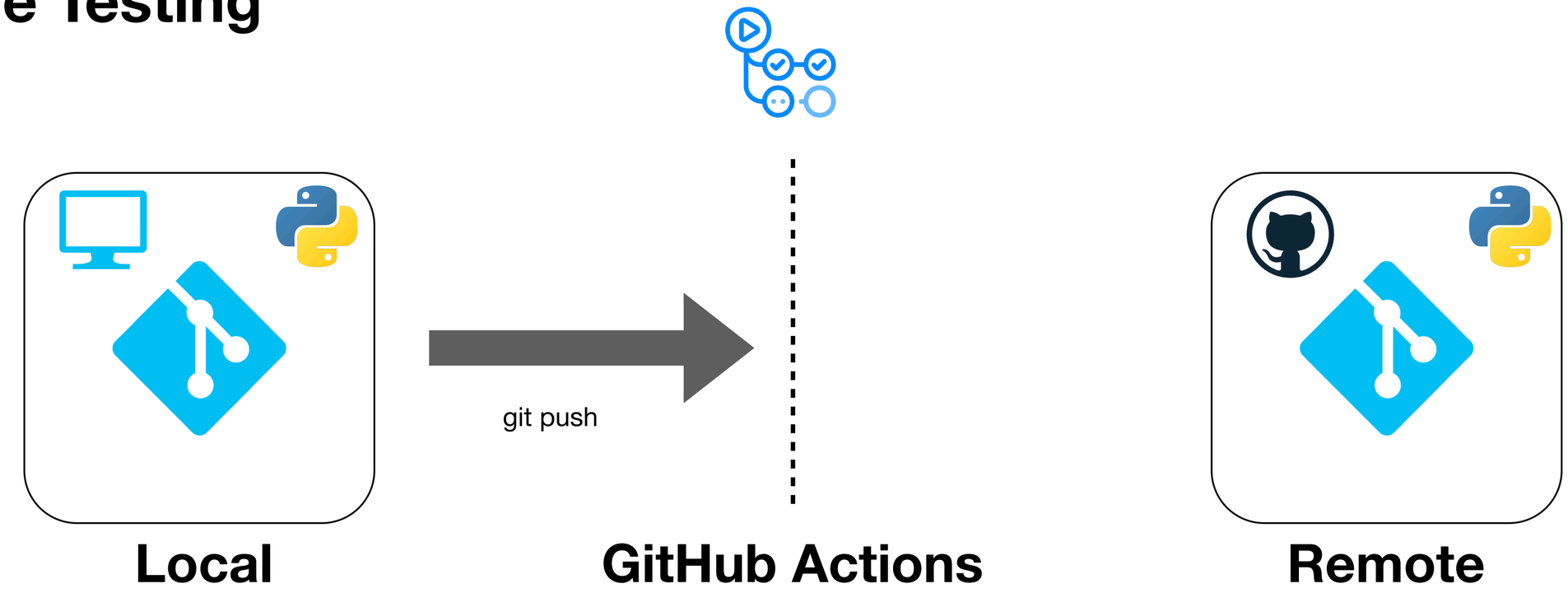


Local

Remote

Triggered Workflow

Code Testing



```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

Source: GitHub Actions documentation

```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

Source: GitHub Actions documentation

```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

Source: GitHub Actions documentation

```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

Source: GitHub Actions documentation

```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

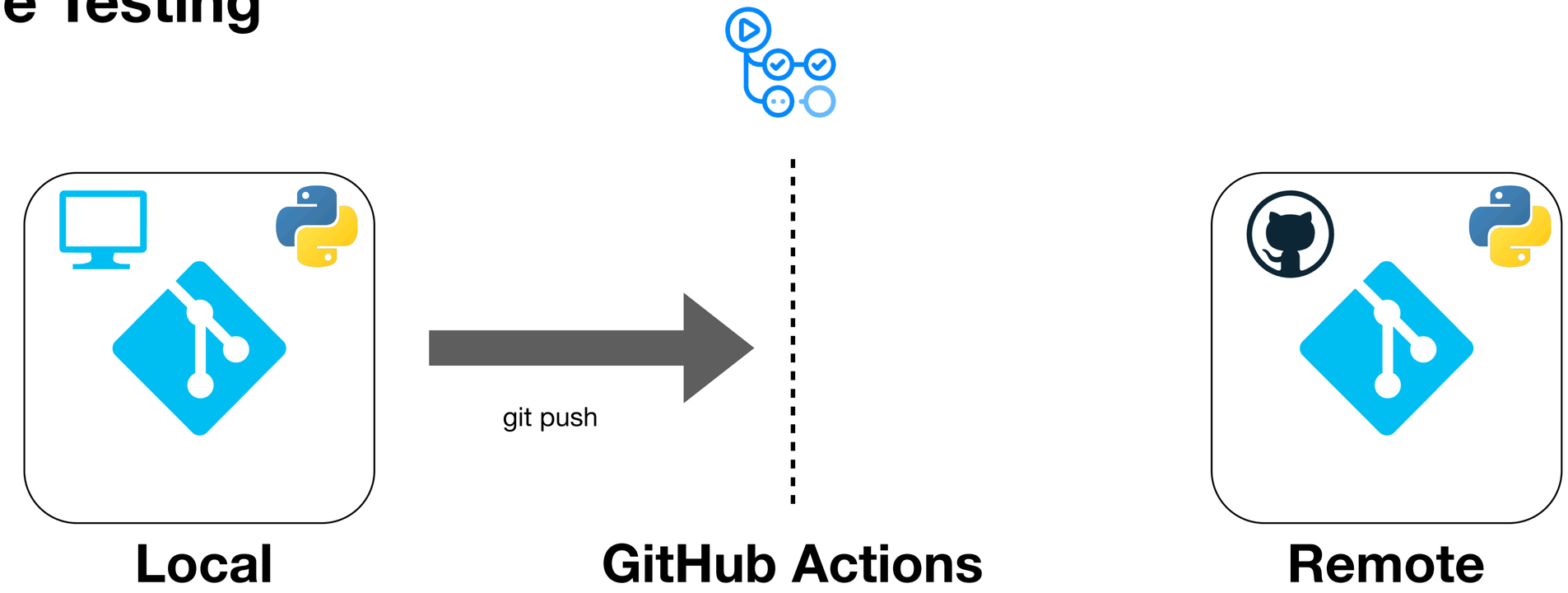
Source: GitHub Actions documentation

```yaml
name: Python package

on: [push]

jobs:
  build:

    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: [ubuntu-latest, macos-latest, windows-latest]
        python-version: ["3.9", "3.11", "3.13", "pypy3.10"]
        exclude:
          - os: macos-latest
            python-version: "3.11"
          - os: windows-latest
            python-version: "3.11"
```

Source: GitHub Actions documentation

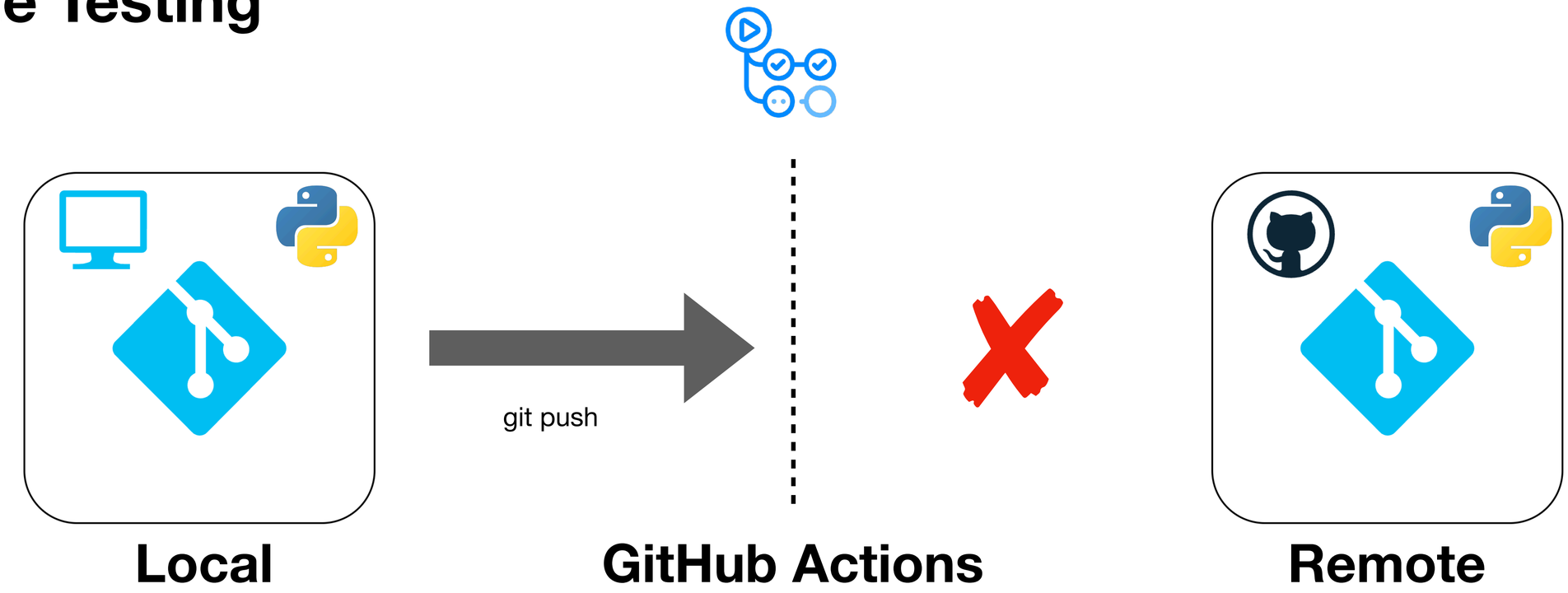
Triggered Workflow

Code Testing



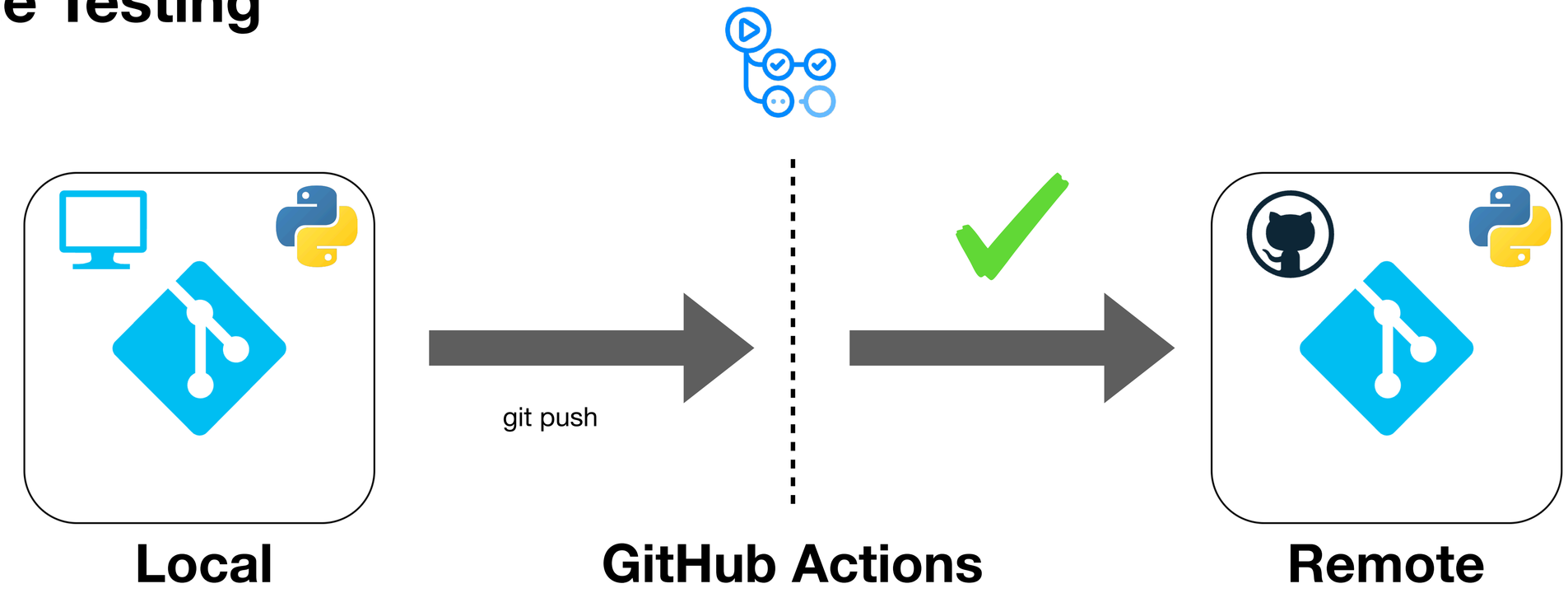
Triggered Workflow

Code Testing



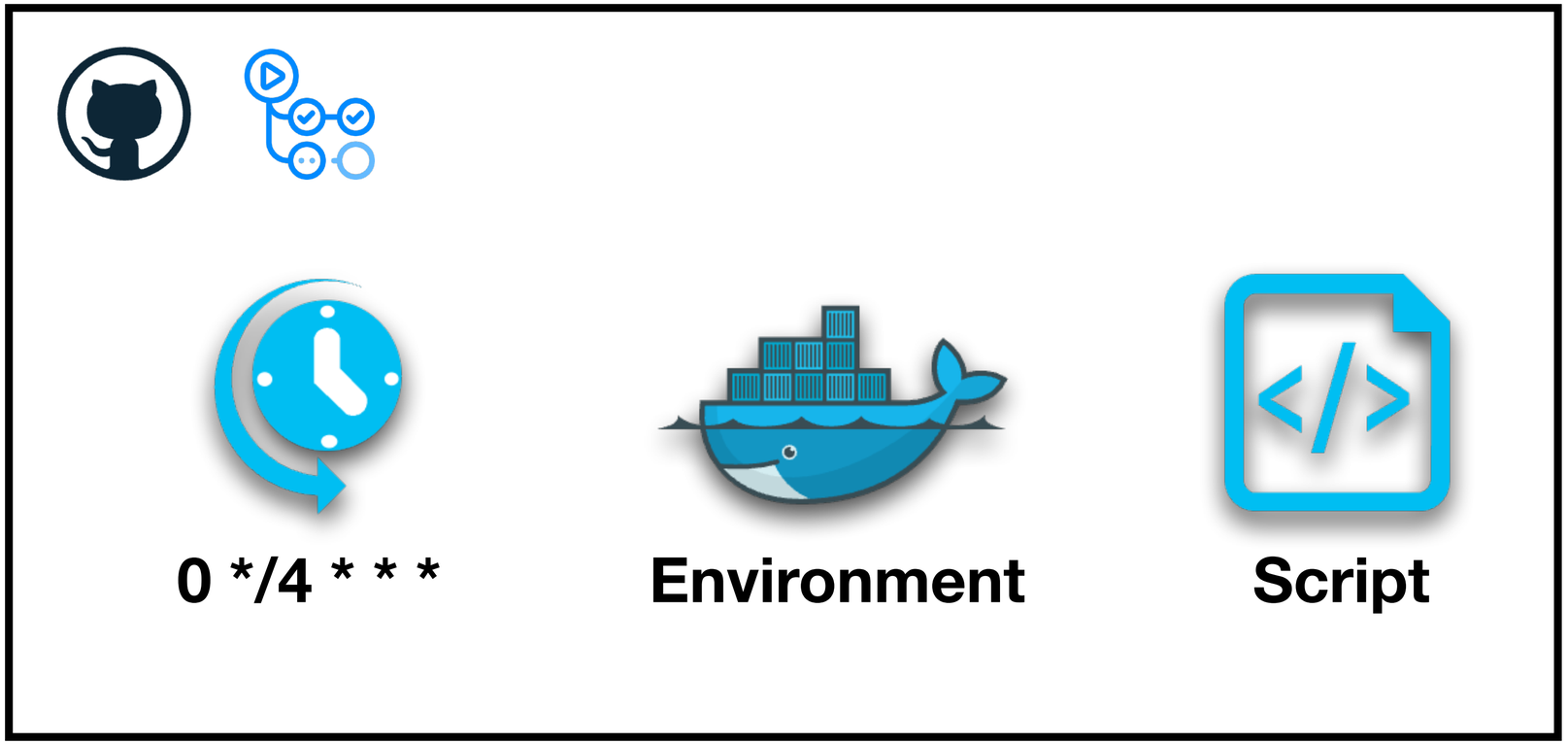
Triggered Workflow

Code Testing



Scheduled Workflow

Cron Job



GitHub Actions

Pros vs. Cons

- Out of the box
- Scale
- Ecosystem

GitHub Actions

Pros vs. Cons

- Out of the box
- Scale
- Ecosystem

- Streaming
- Orchestration
- Debugging

GitHub Actions

Pros vs. Cons

Forecasting

A Live Anomaly Detection

GitHub Actions

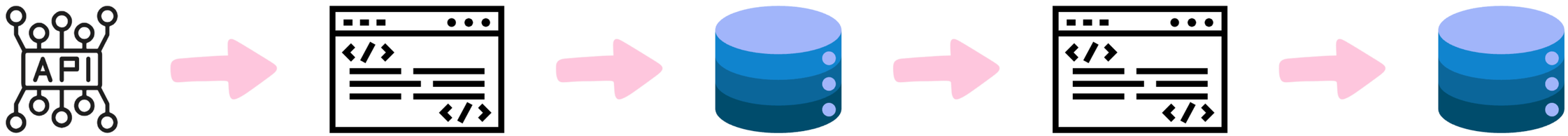
Key Features

- Multiple OS support
- Docker support
- Any language
- Secrets
- Logs

Data & ML Pipelines with Actions

In a Nutshell

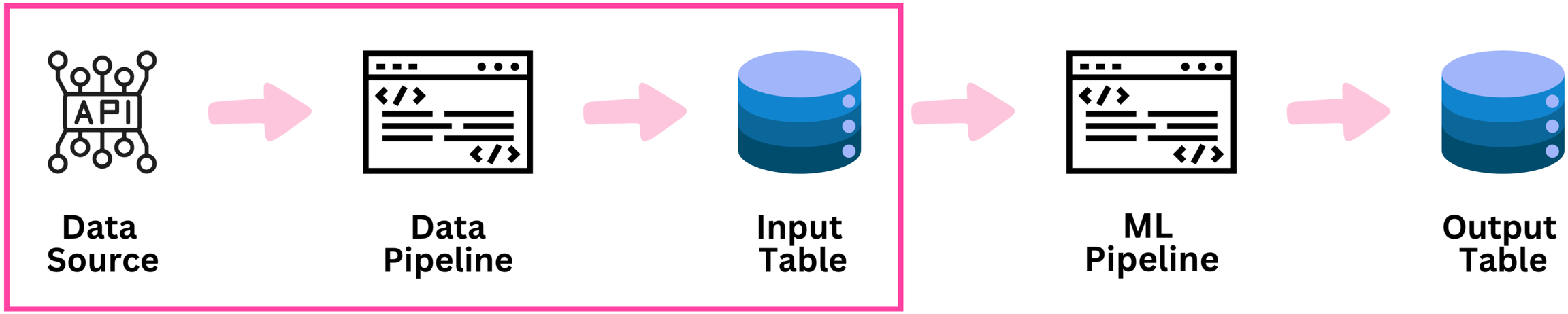
The Pipeline Components



Data Source → Data Pipeline → Input Table → ML Pipeline → Output Table

In a Nutshell

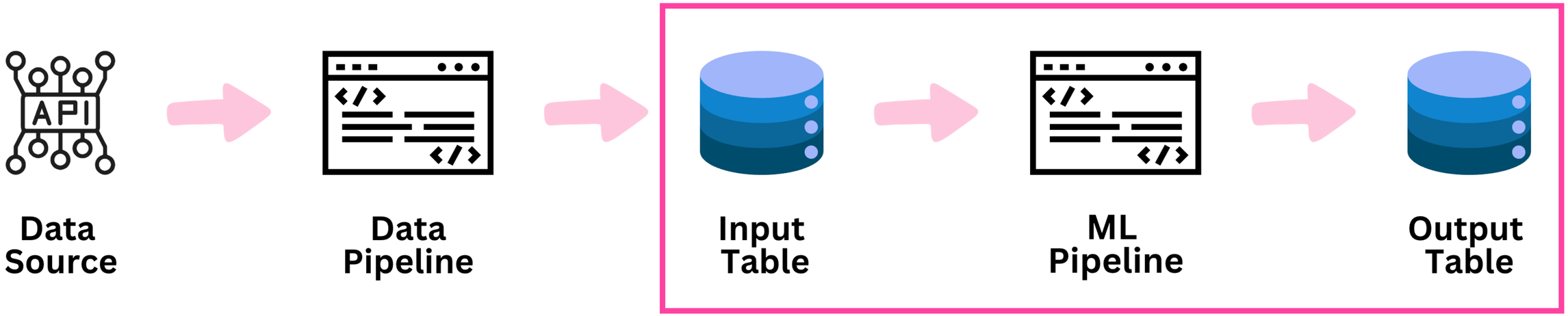
The Pipeline Components



- ETL
- Feature engineering

In a Nutshell

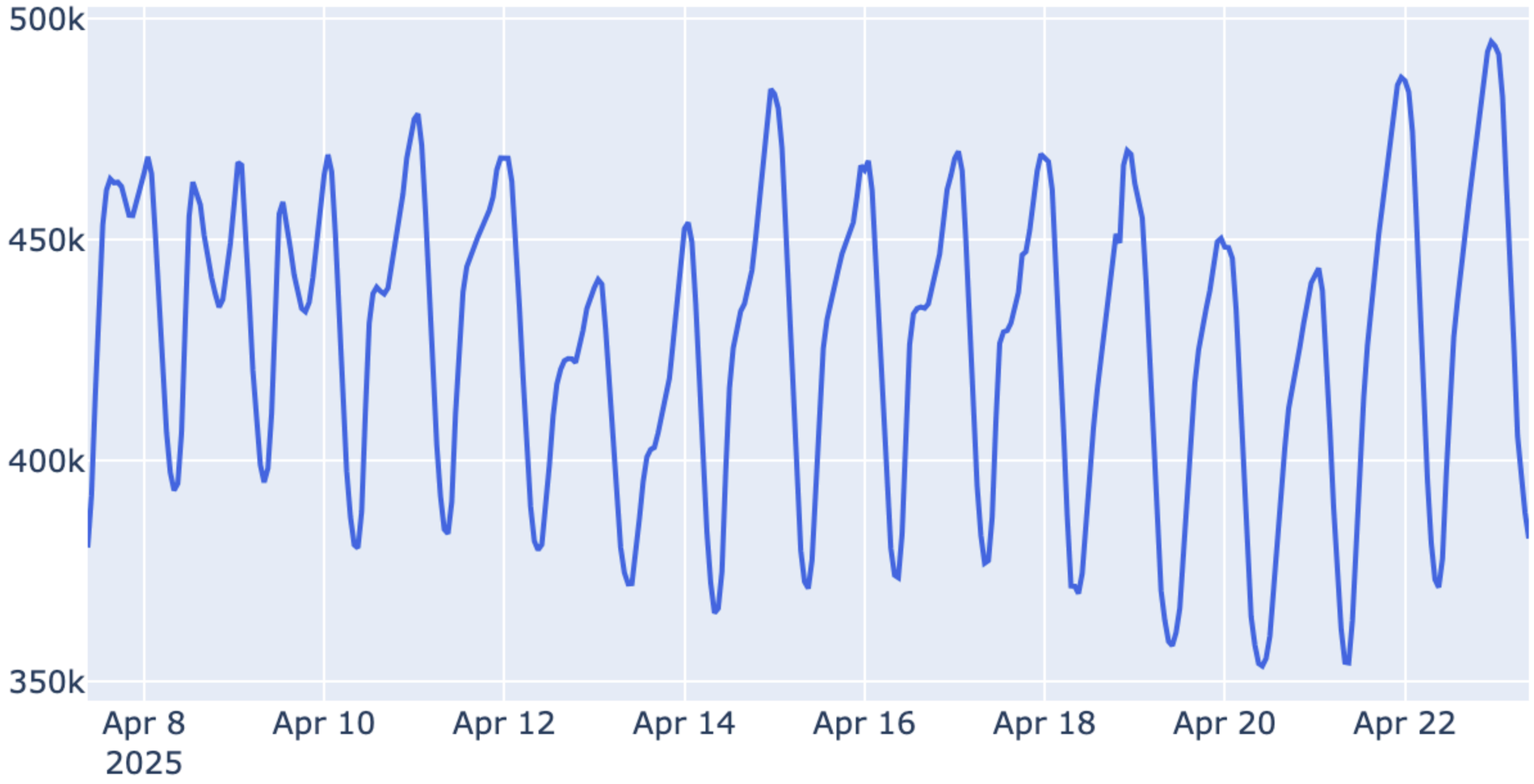
The Pipeline Components



- Refresh the model
- Monitor

Forecasting the US Hourly Demand for Electricity

Forecasting the US Hourly Demand for Electricity



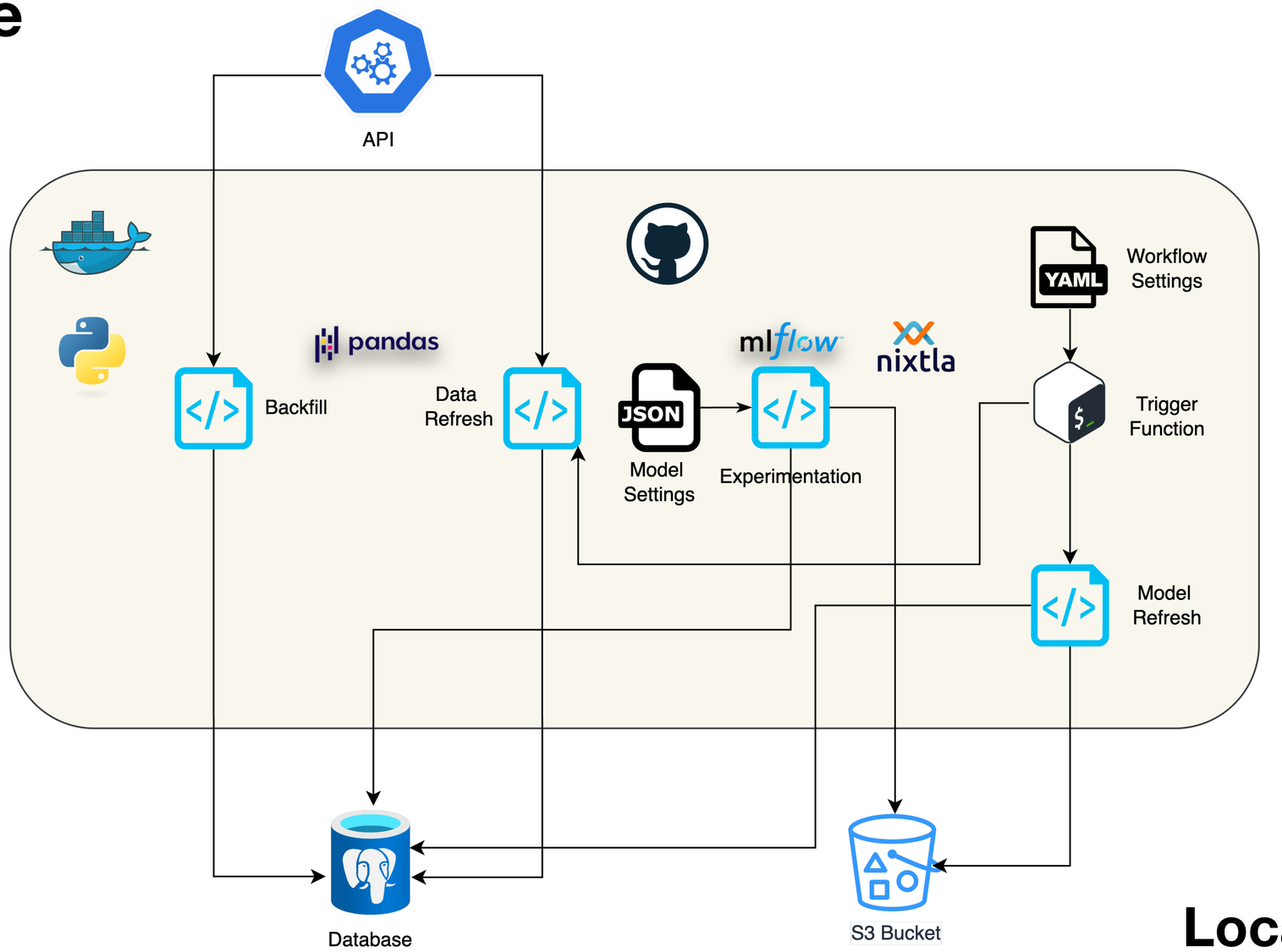
Forecasting the US Hourly Demand for Electricity

Scope

- Refresh the data hourly
- Create 72 hours forecast
- Refresh the forecast every 24 hours

Data & ML Pipeline with GitHub Actions

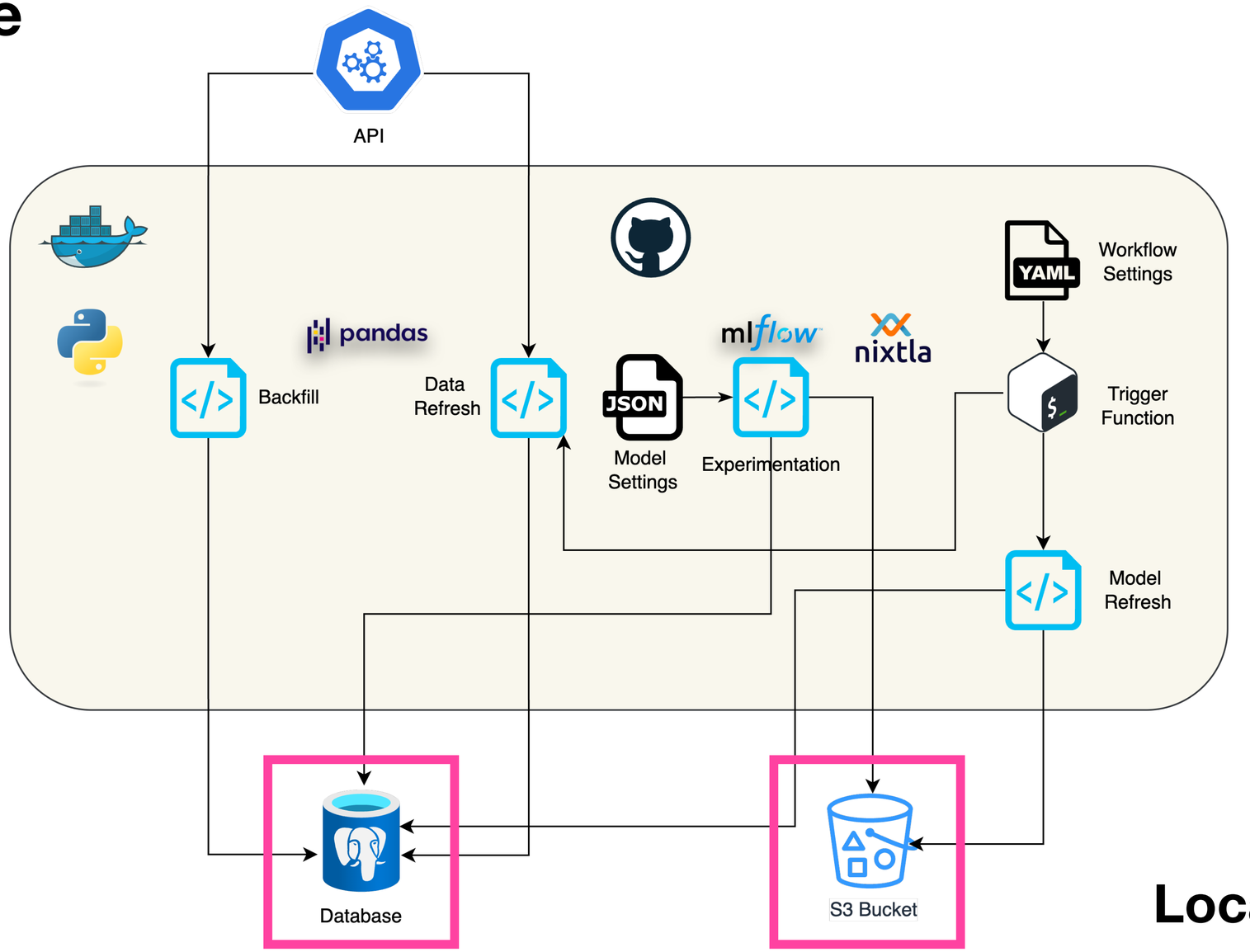
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

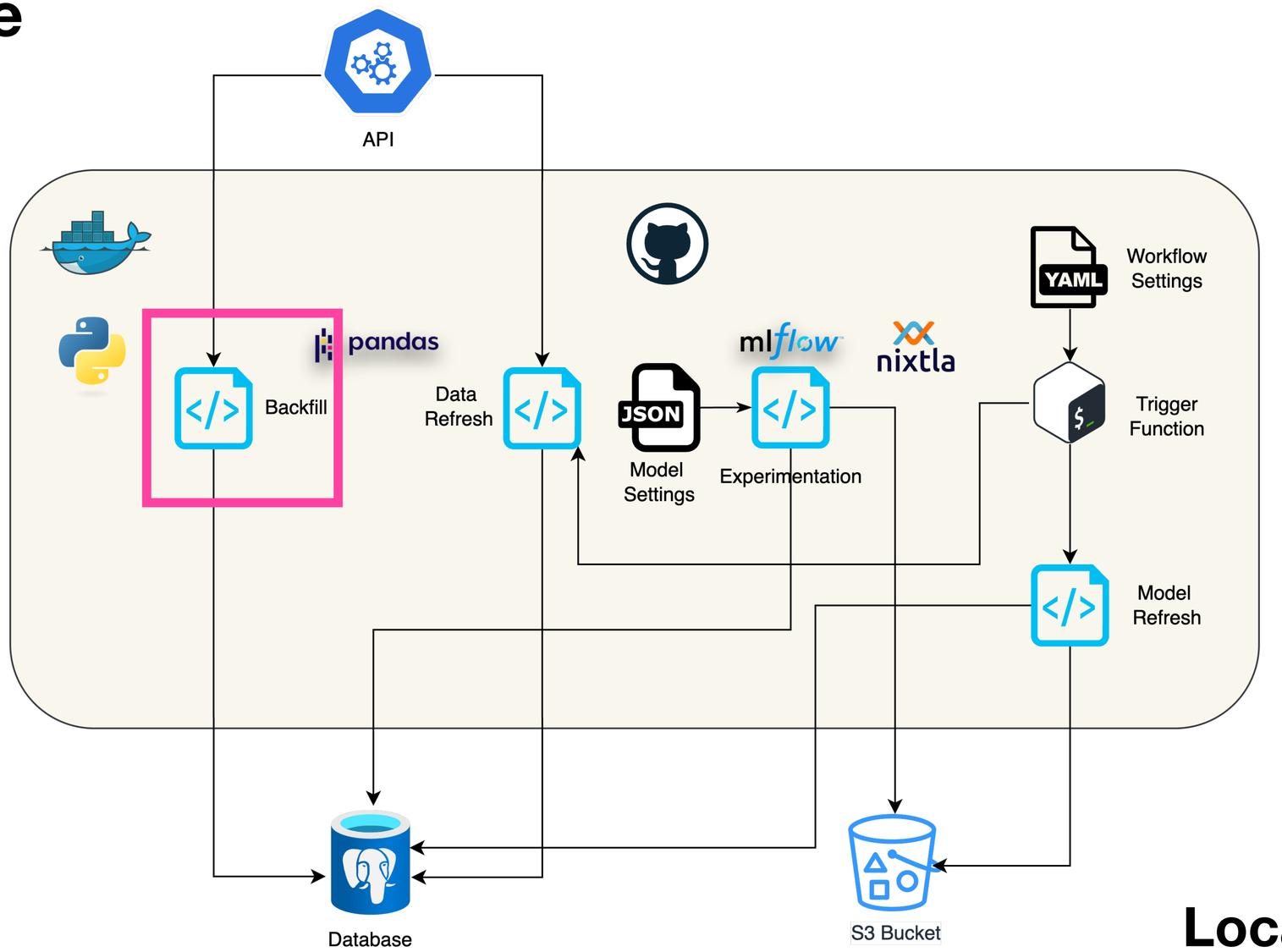
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

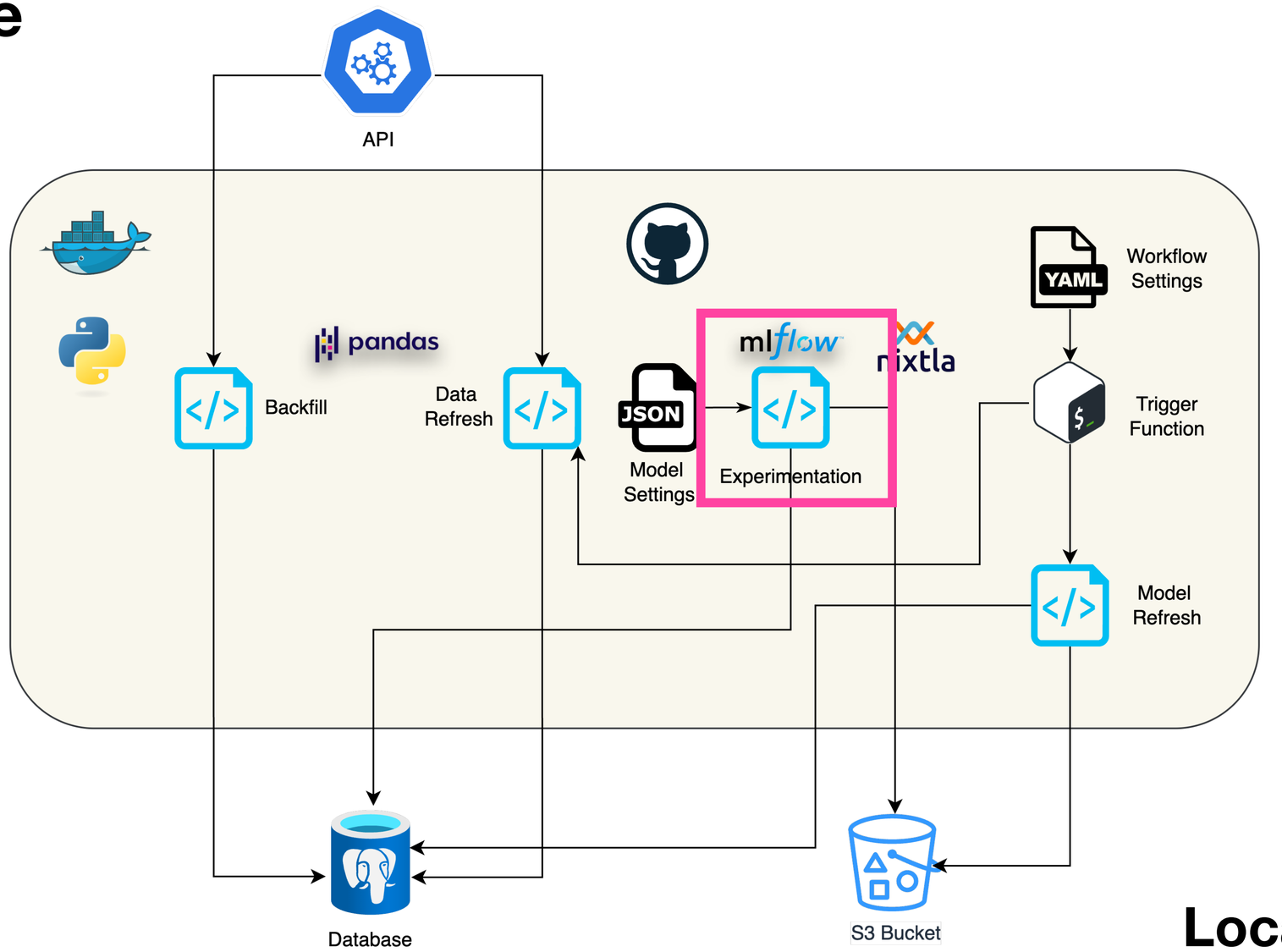
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

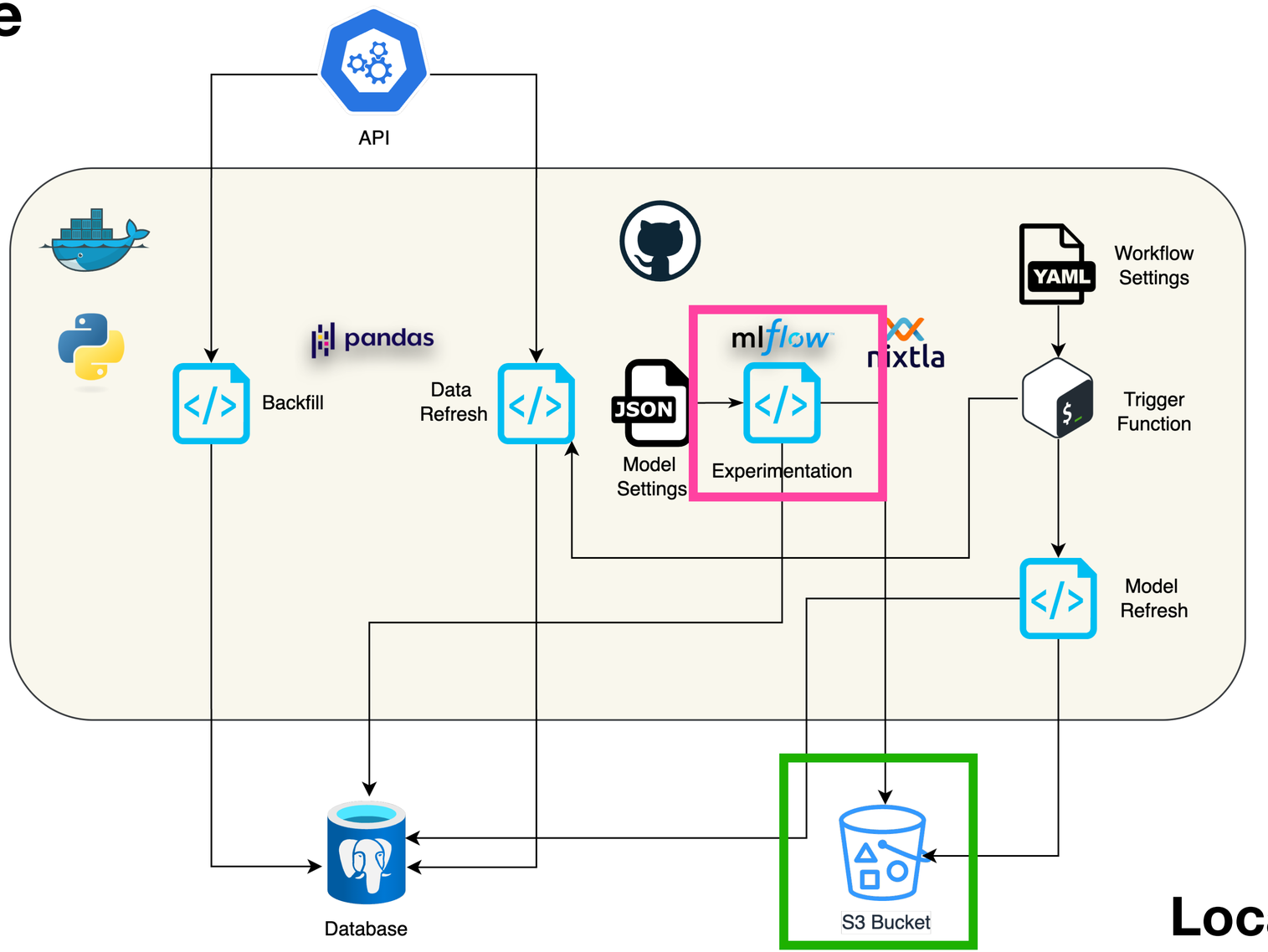
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

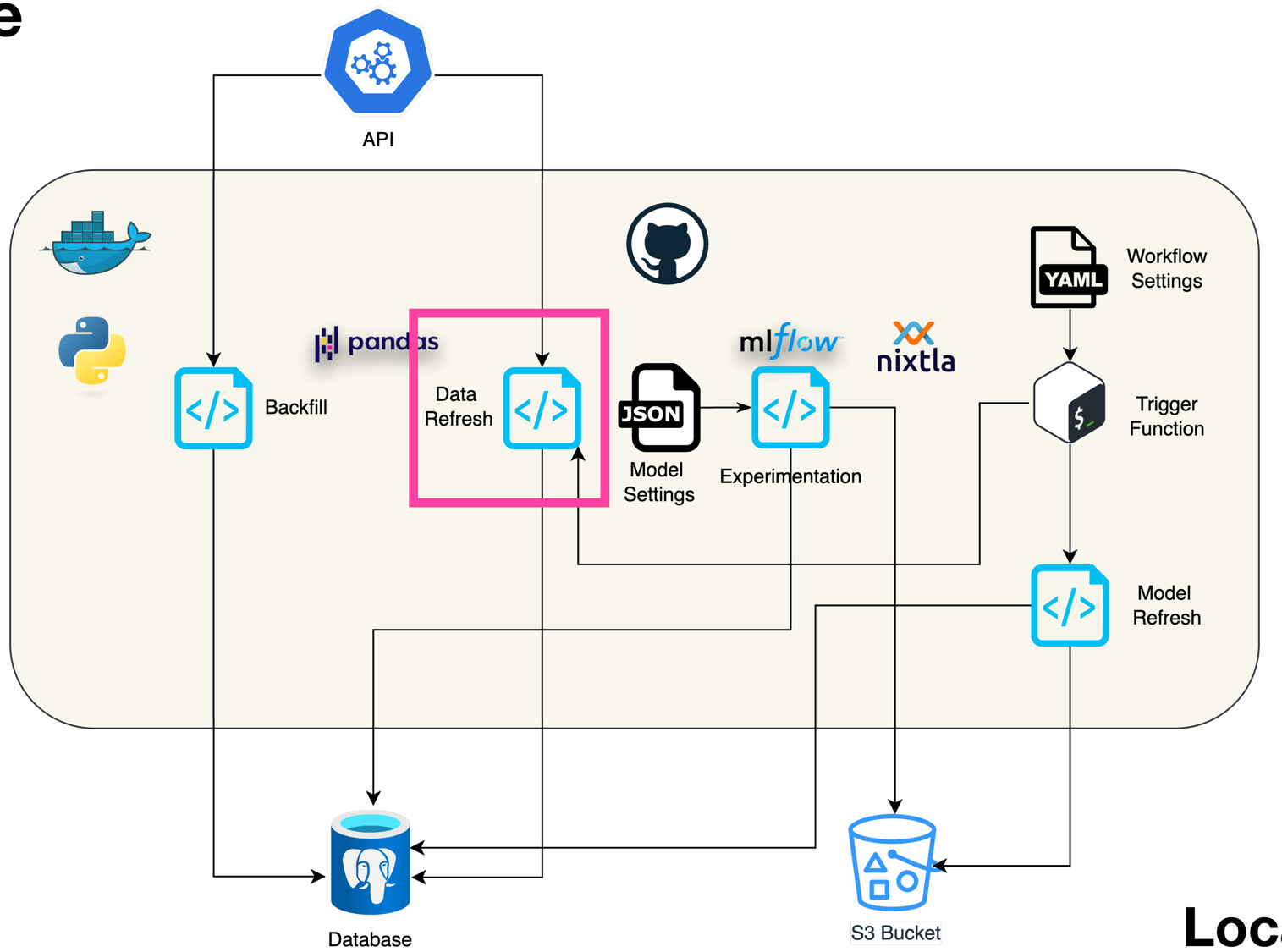
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

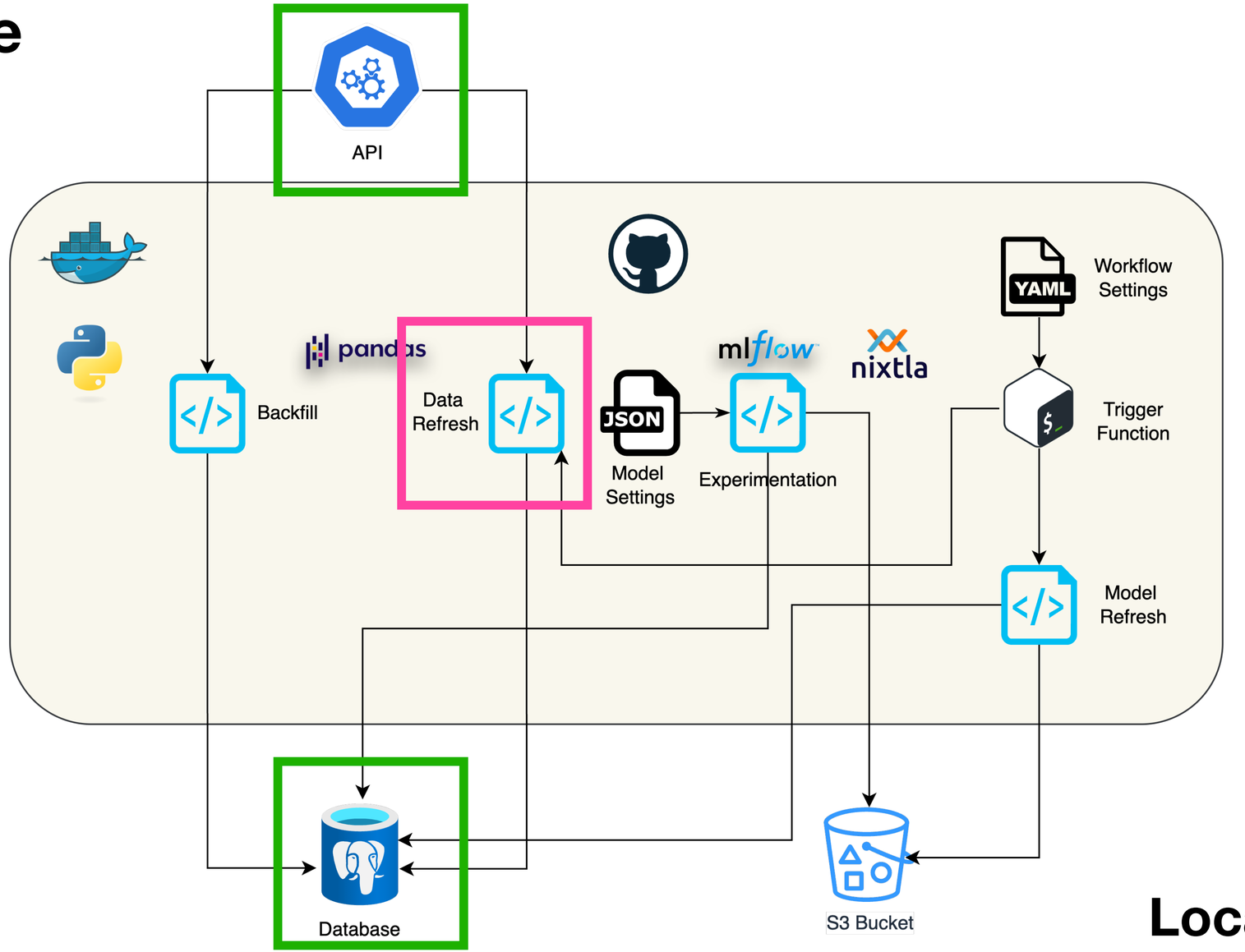
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

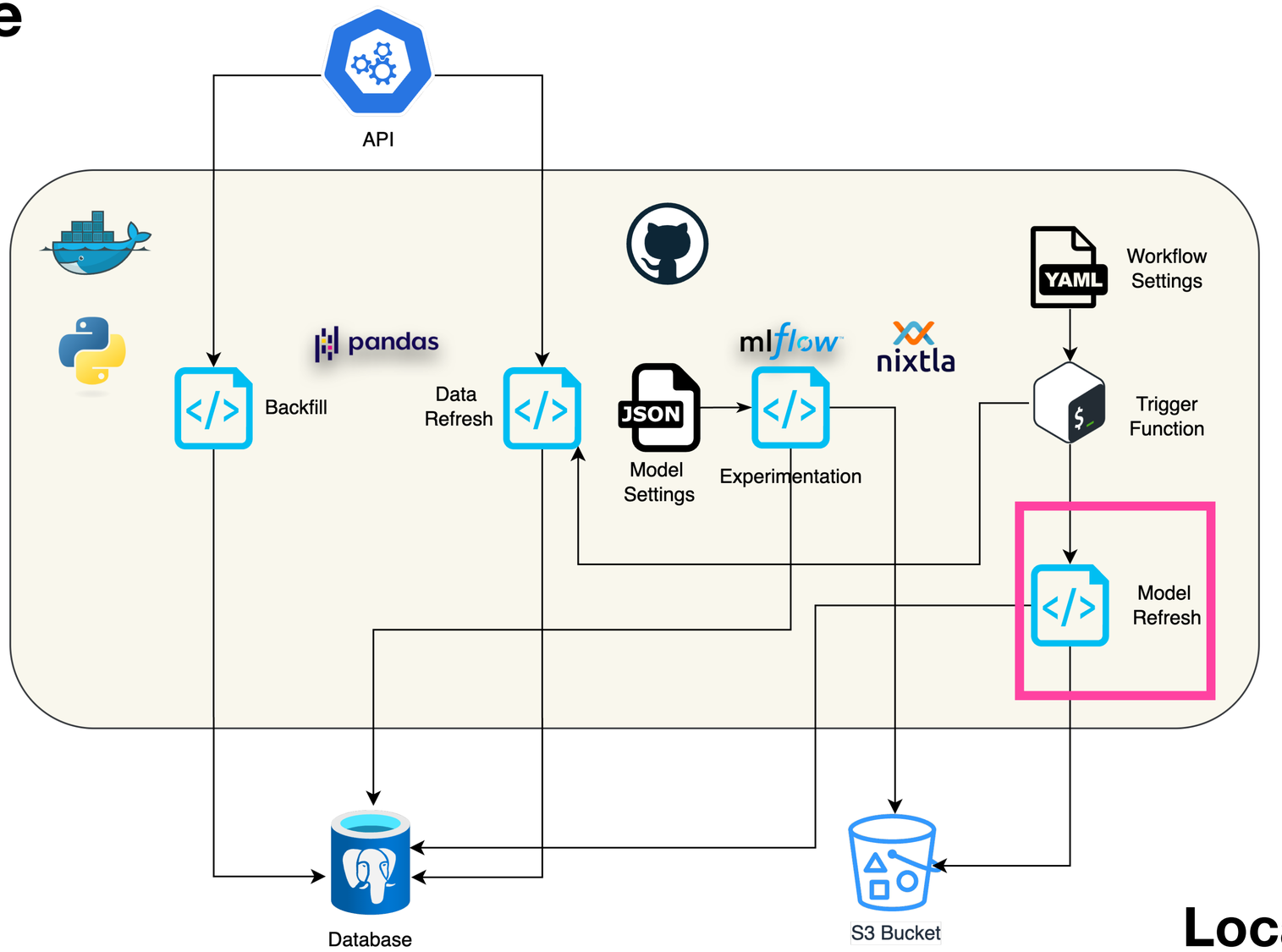
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

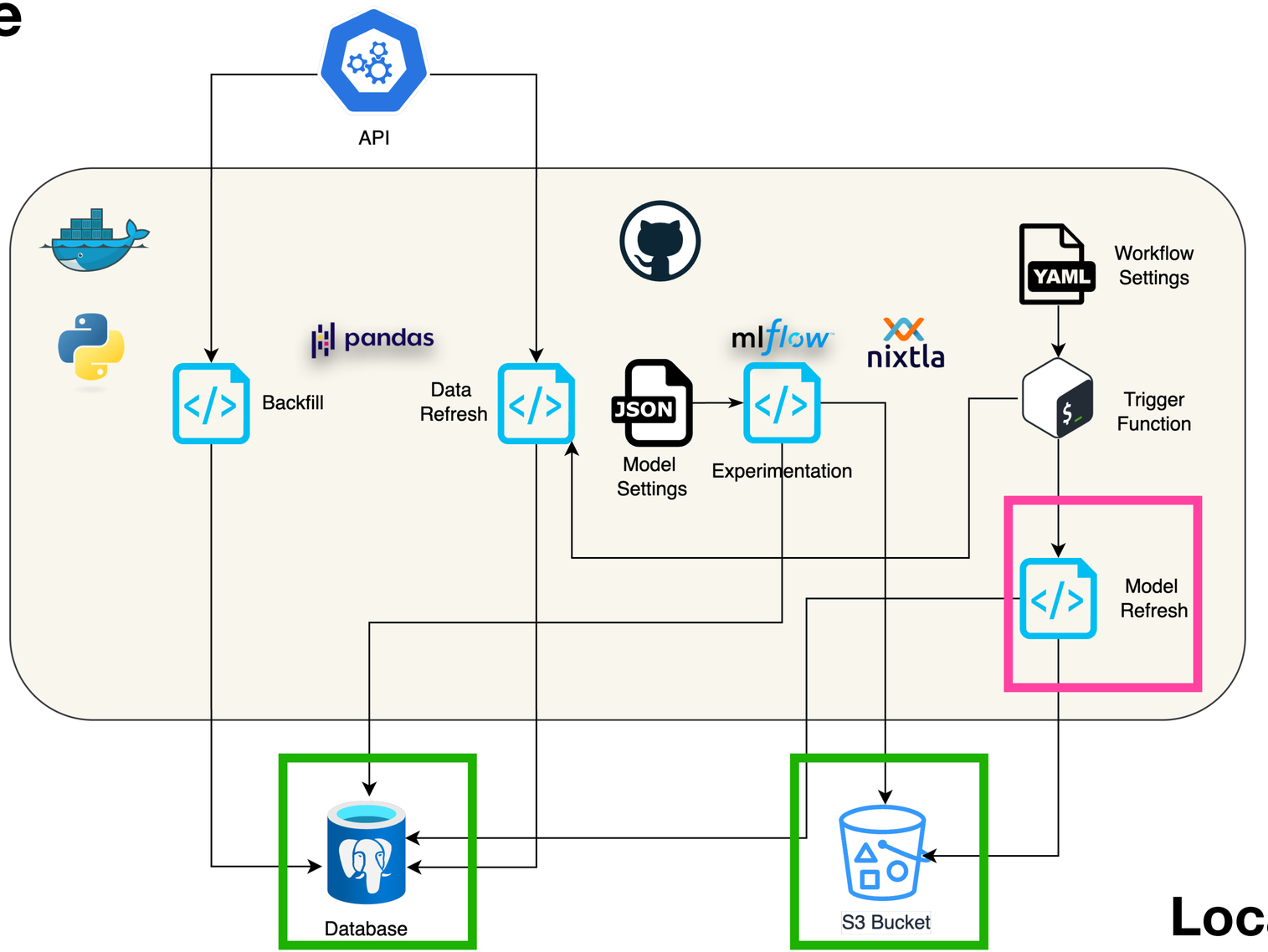
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

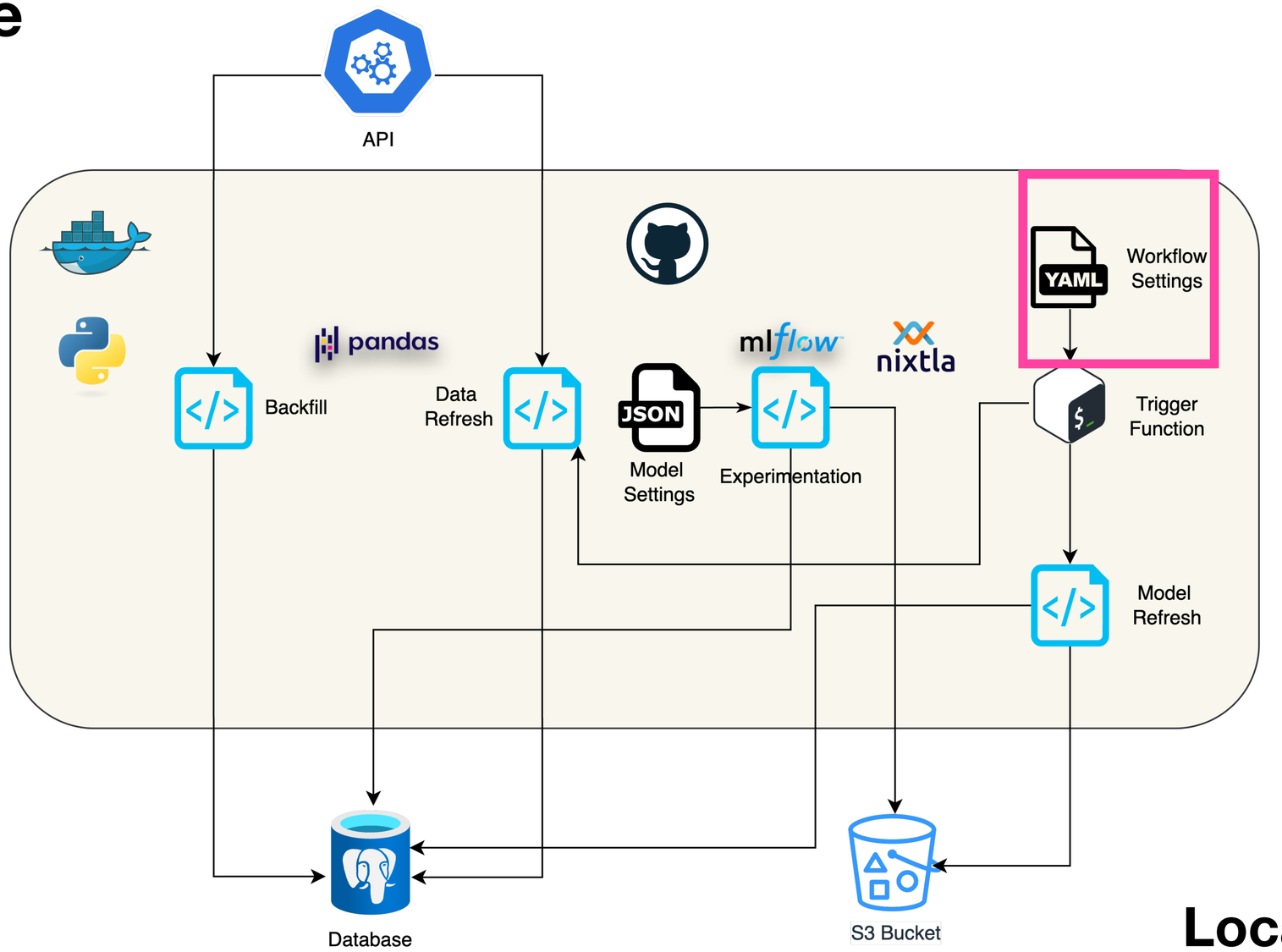
Architecture



Local Repository

Data & ML Pipeline with GitHub Actions

Architecture

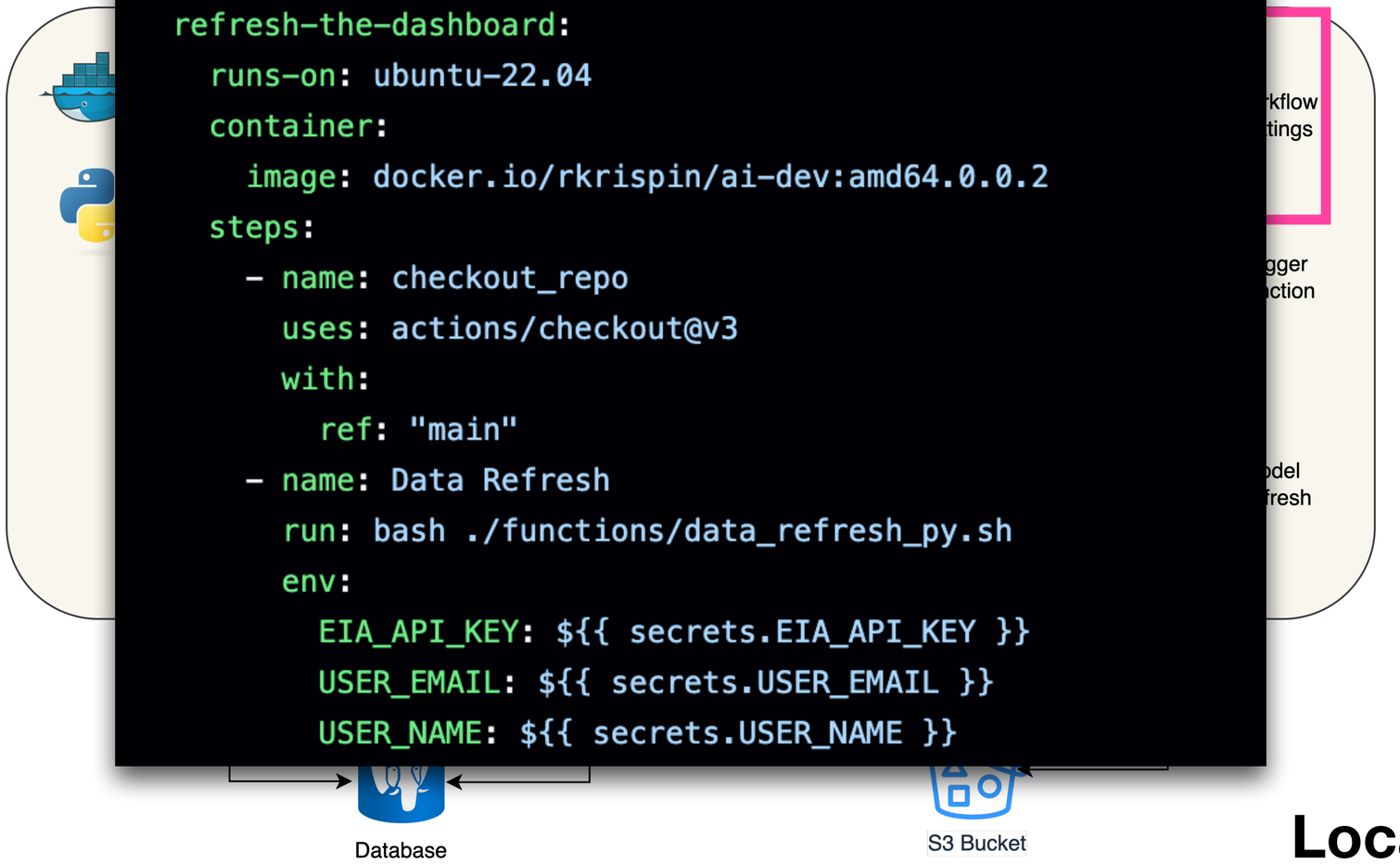


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

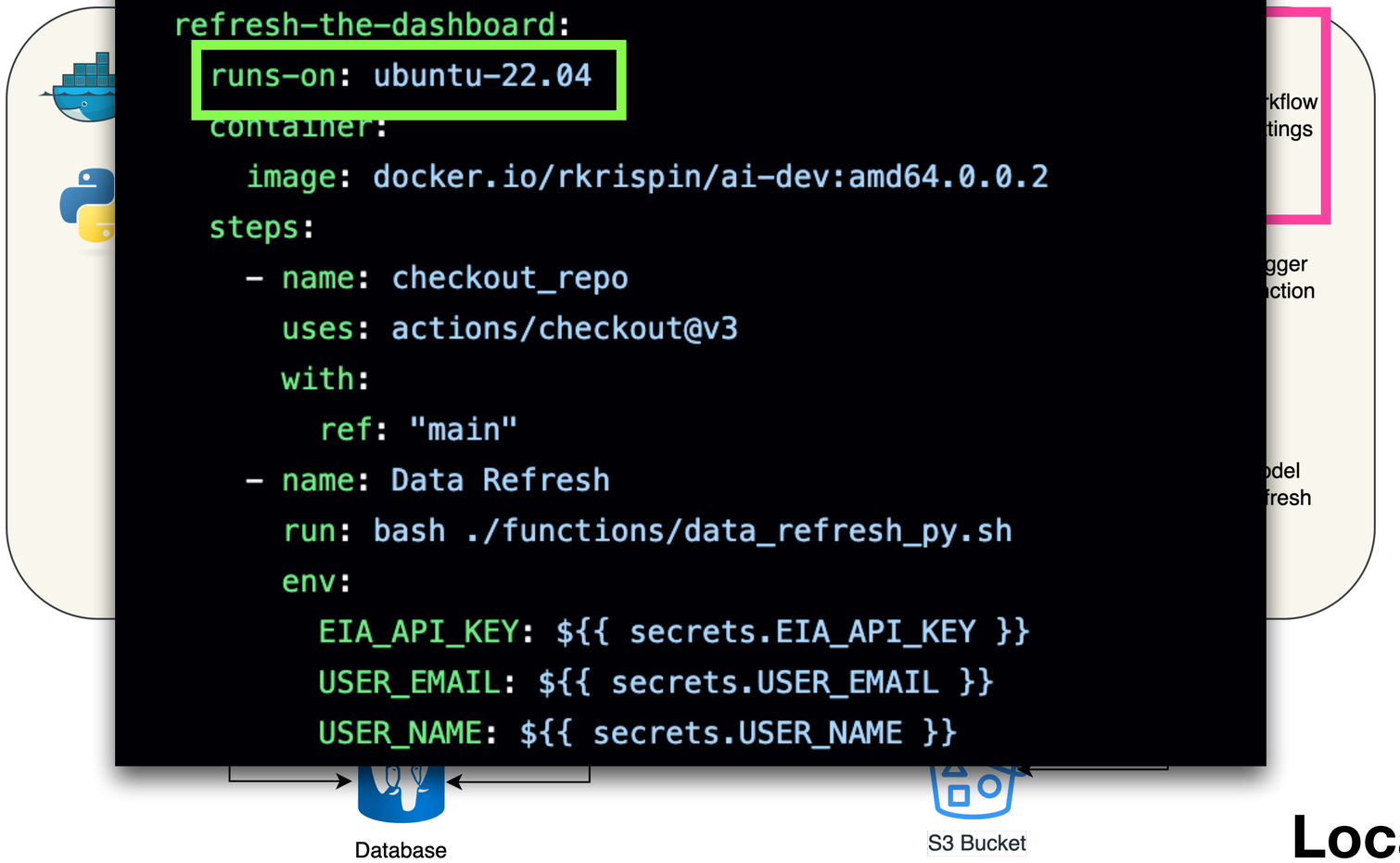


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

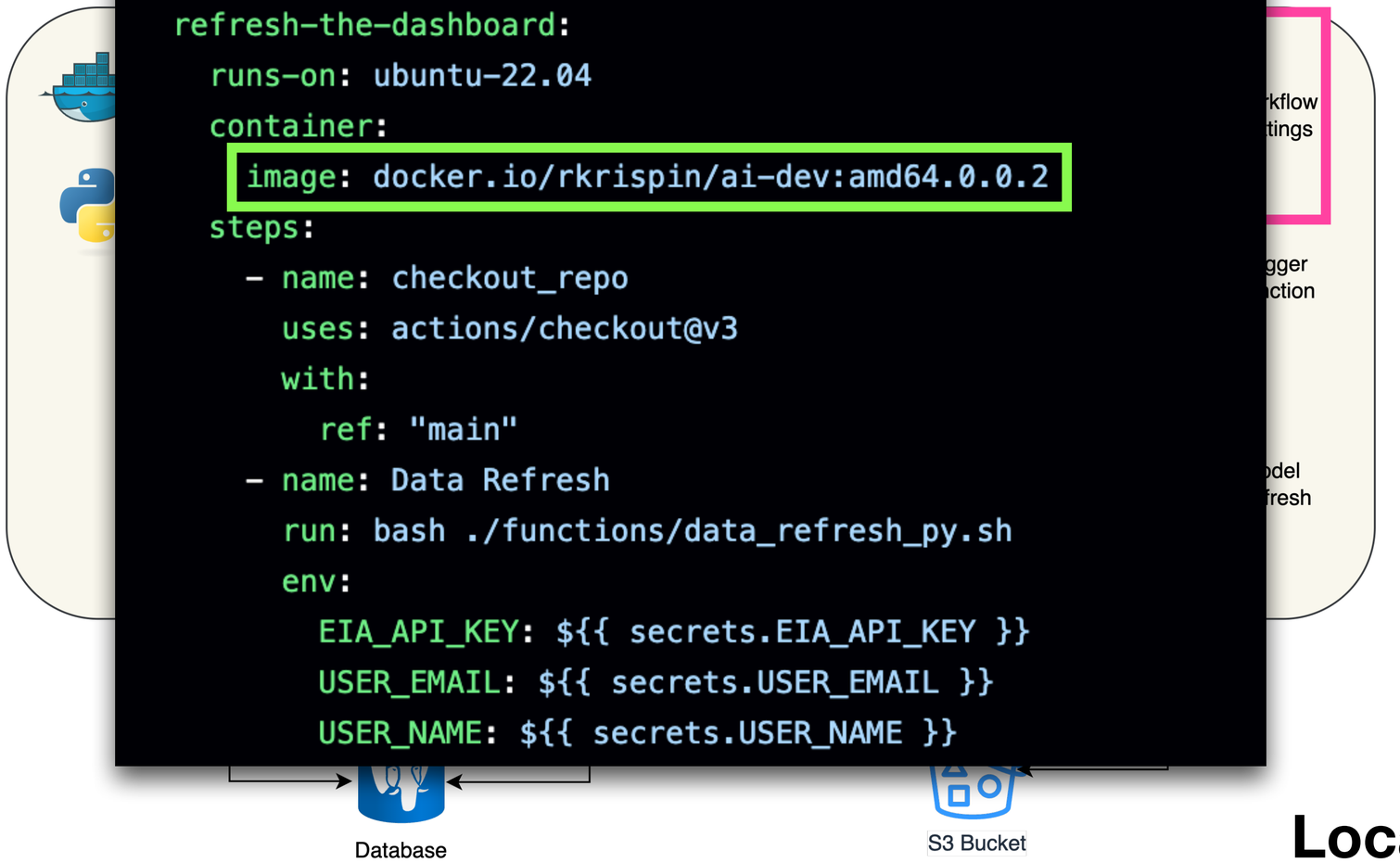


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

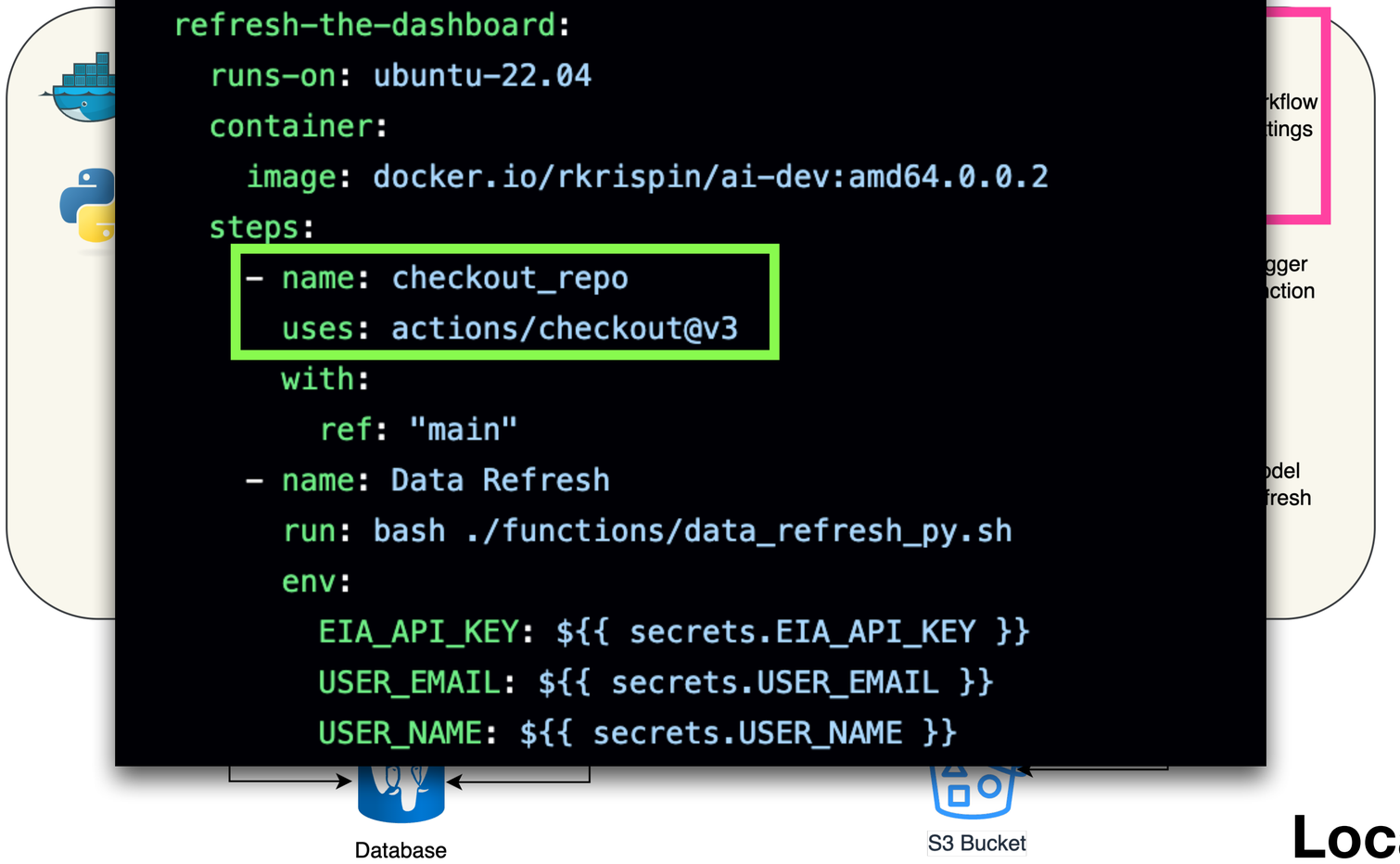


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

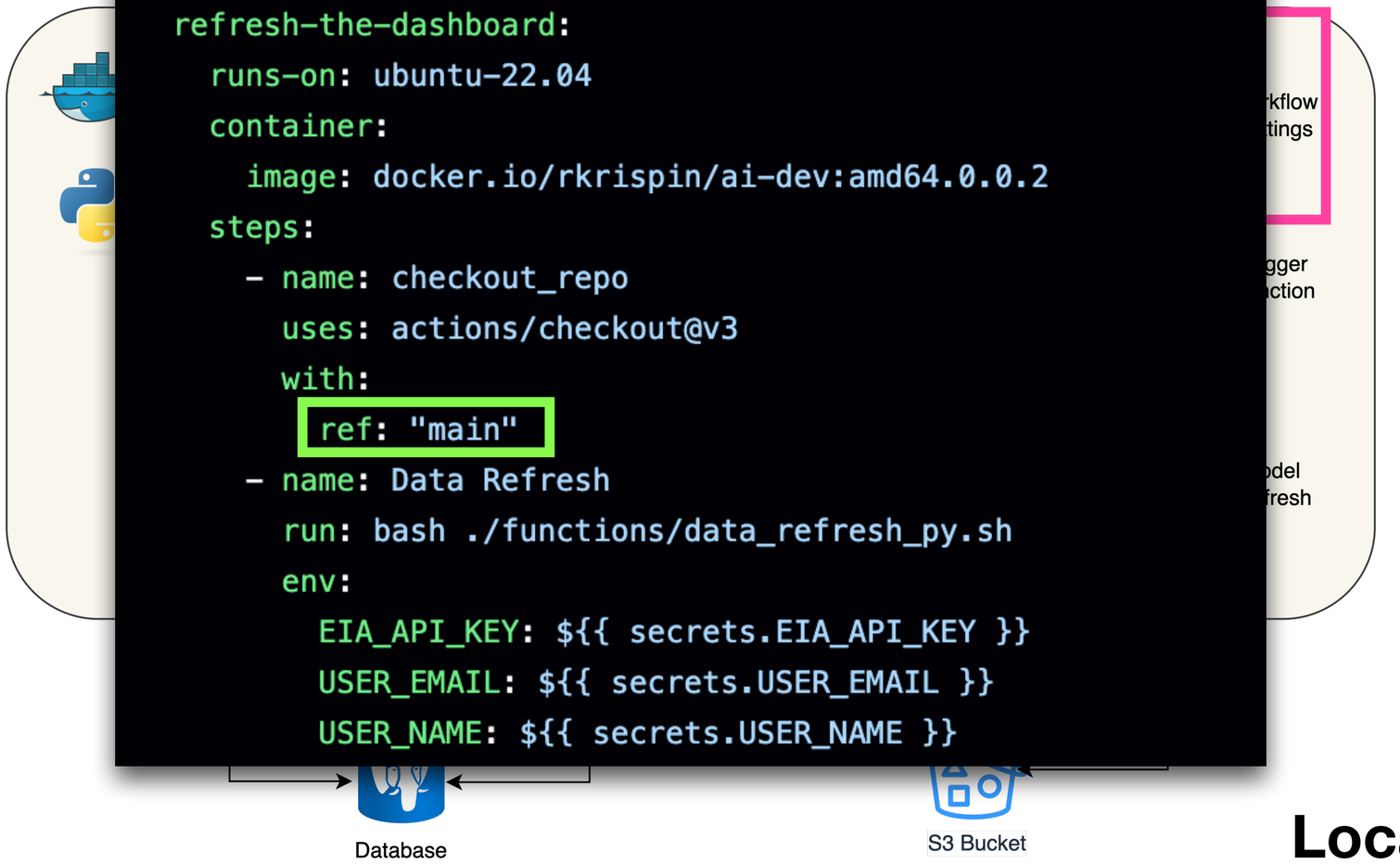


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

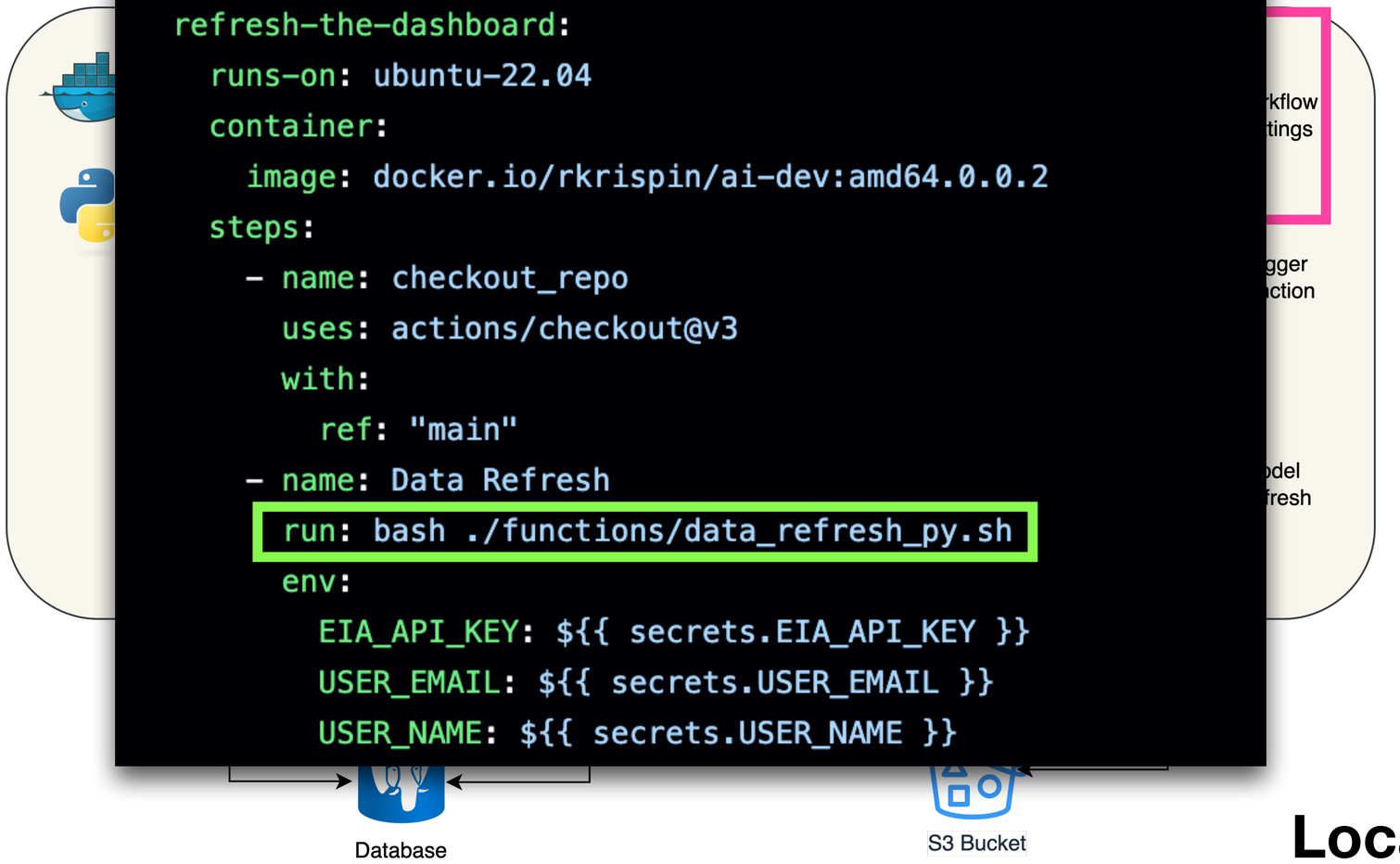


Local Repository

Data & ML ctions

Architecture

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

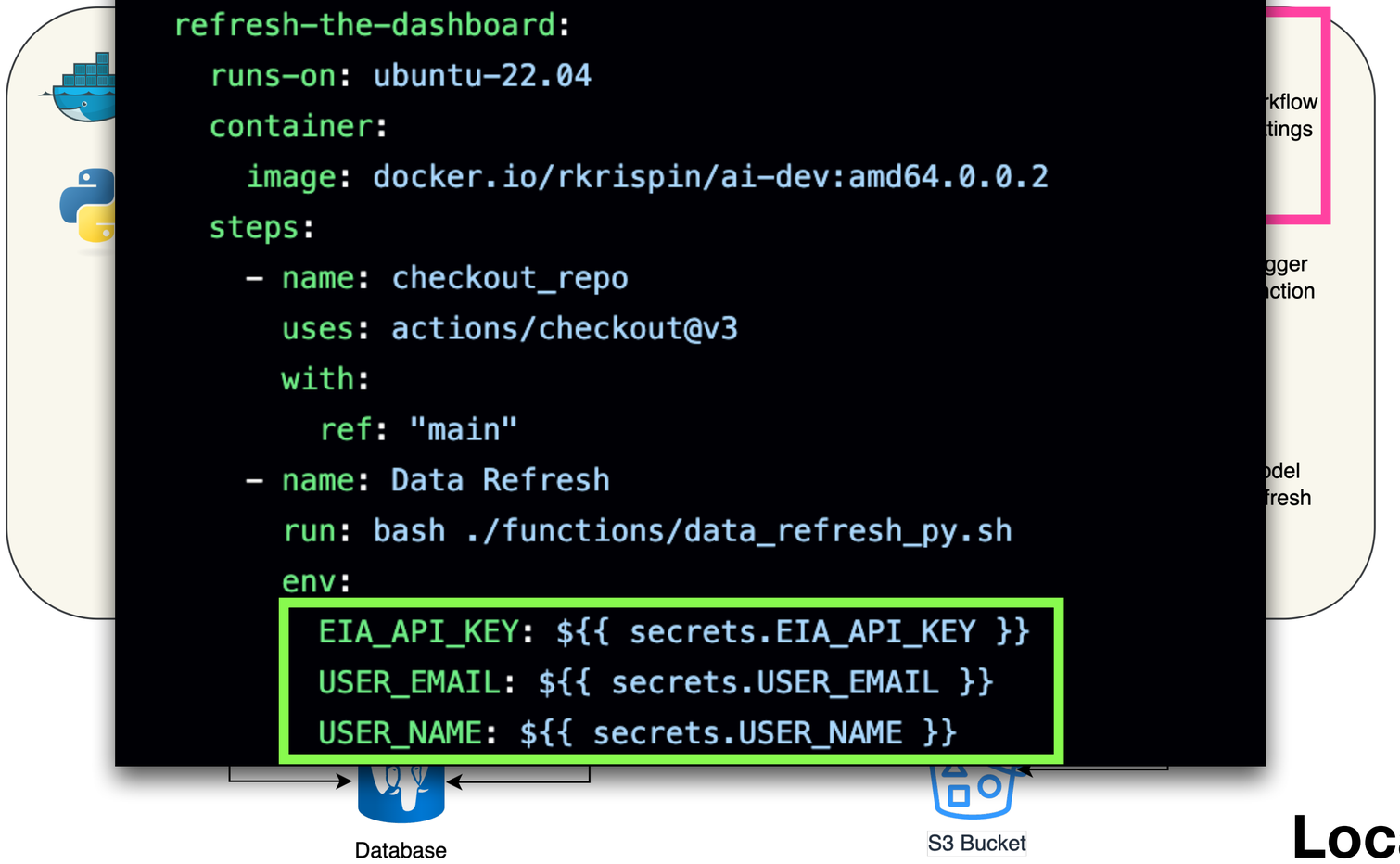


Local Repository

Data & ML ctions

Architecture

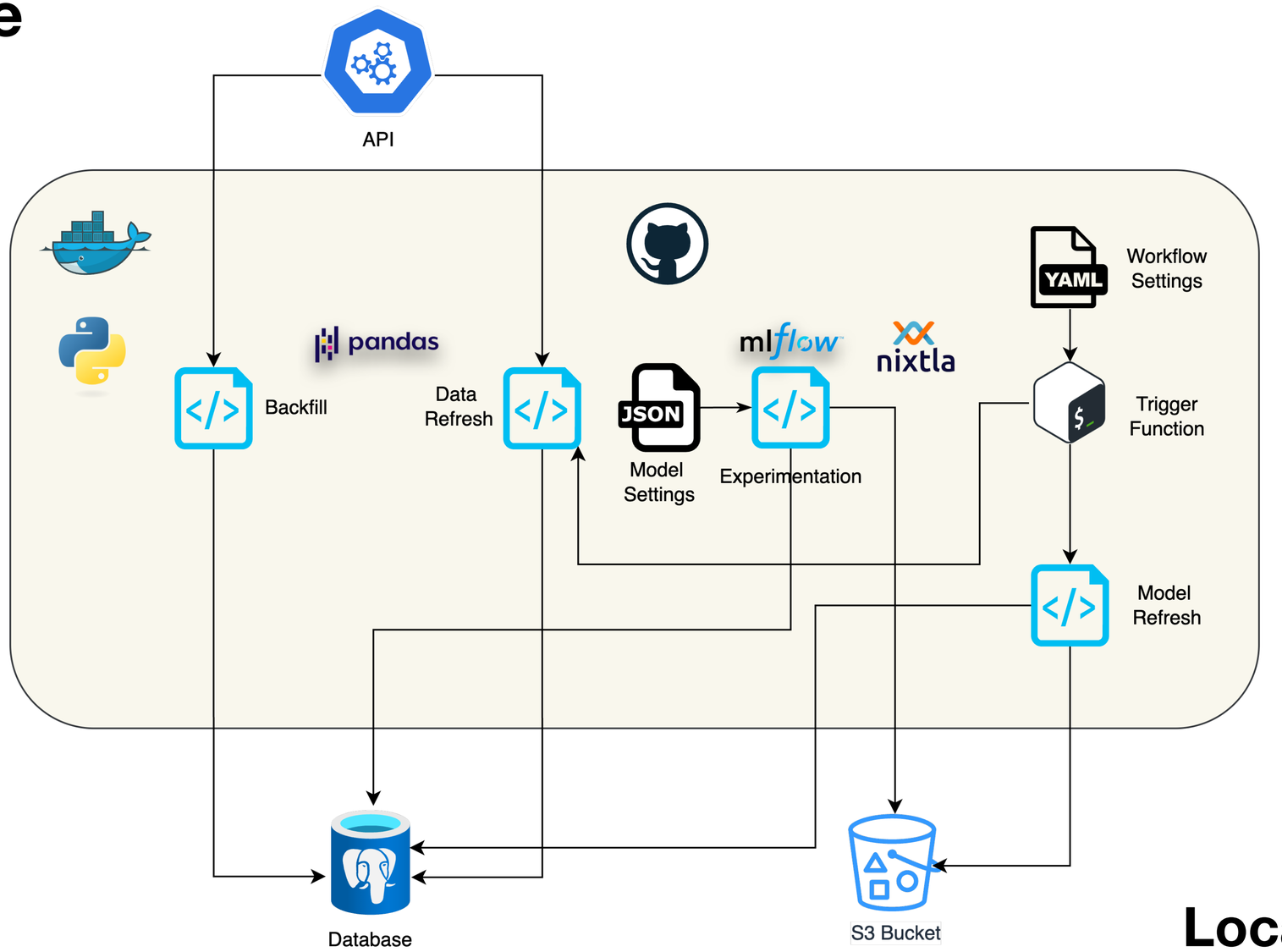
```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */4 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/ai-dev:amd64.0.0.2
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./functions/data_refresh_py.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```



Local Repository

Data & ML Pipeline with GitHub Actions

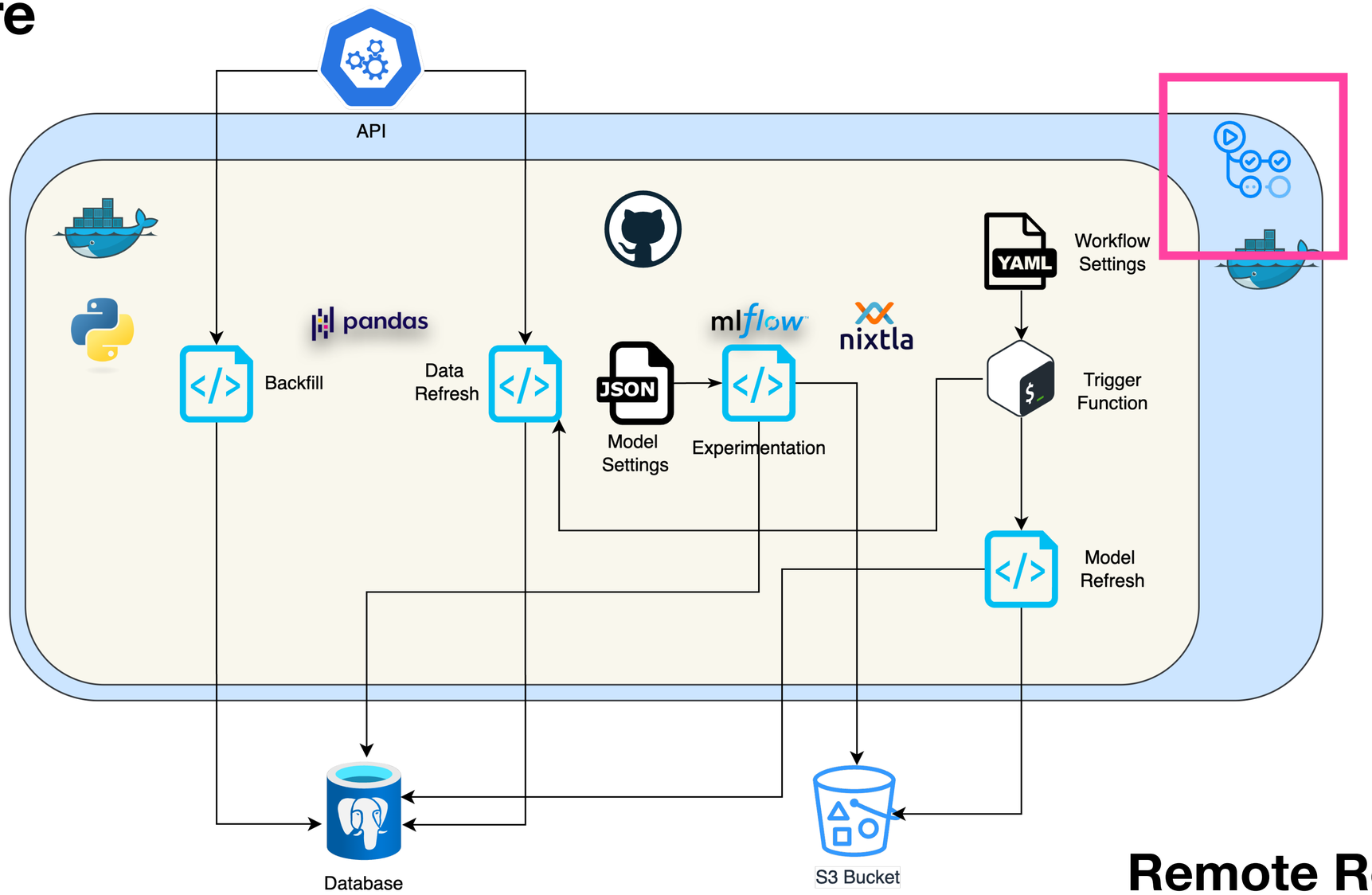
Architecture



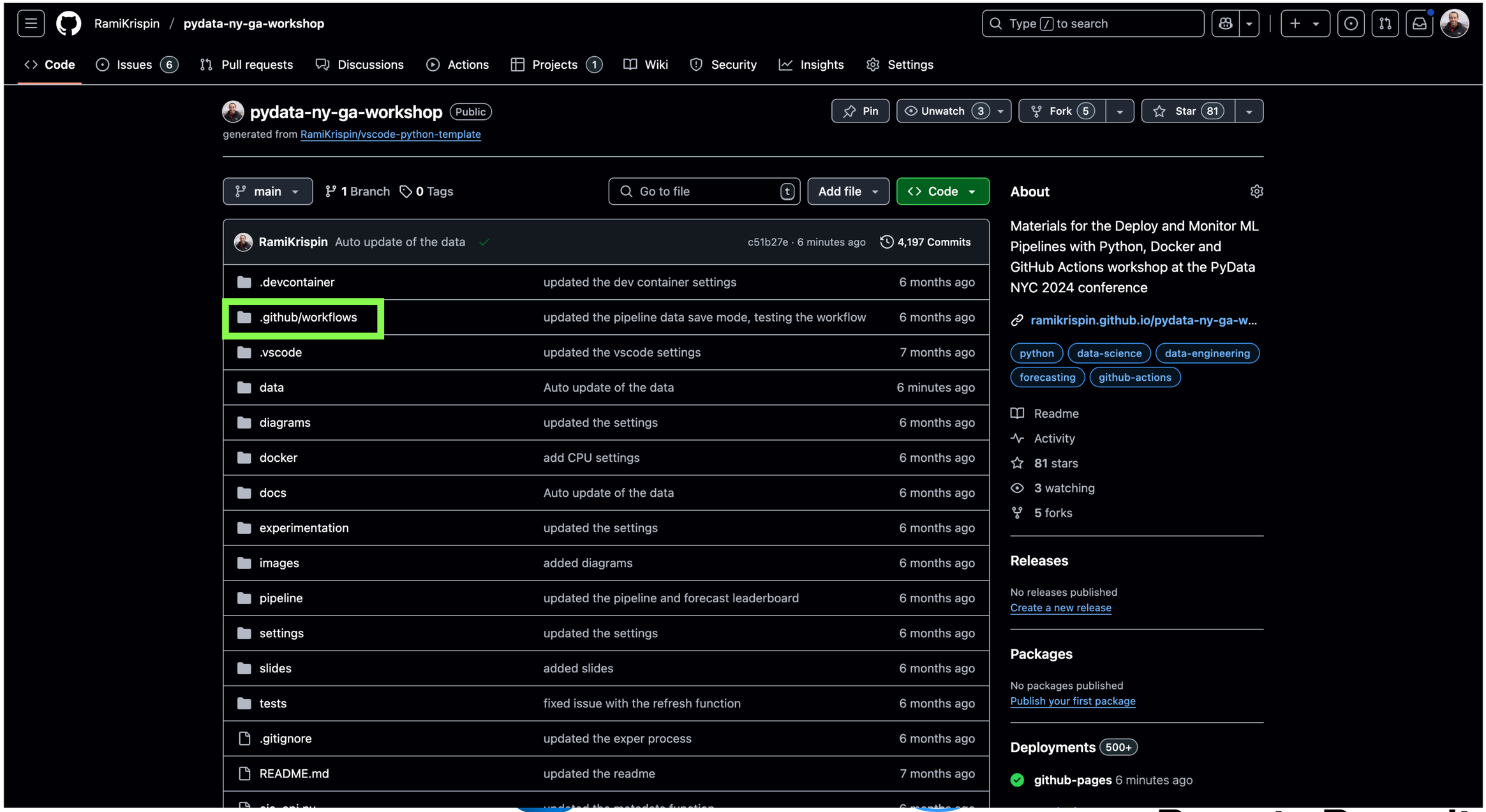
Local Repository

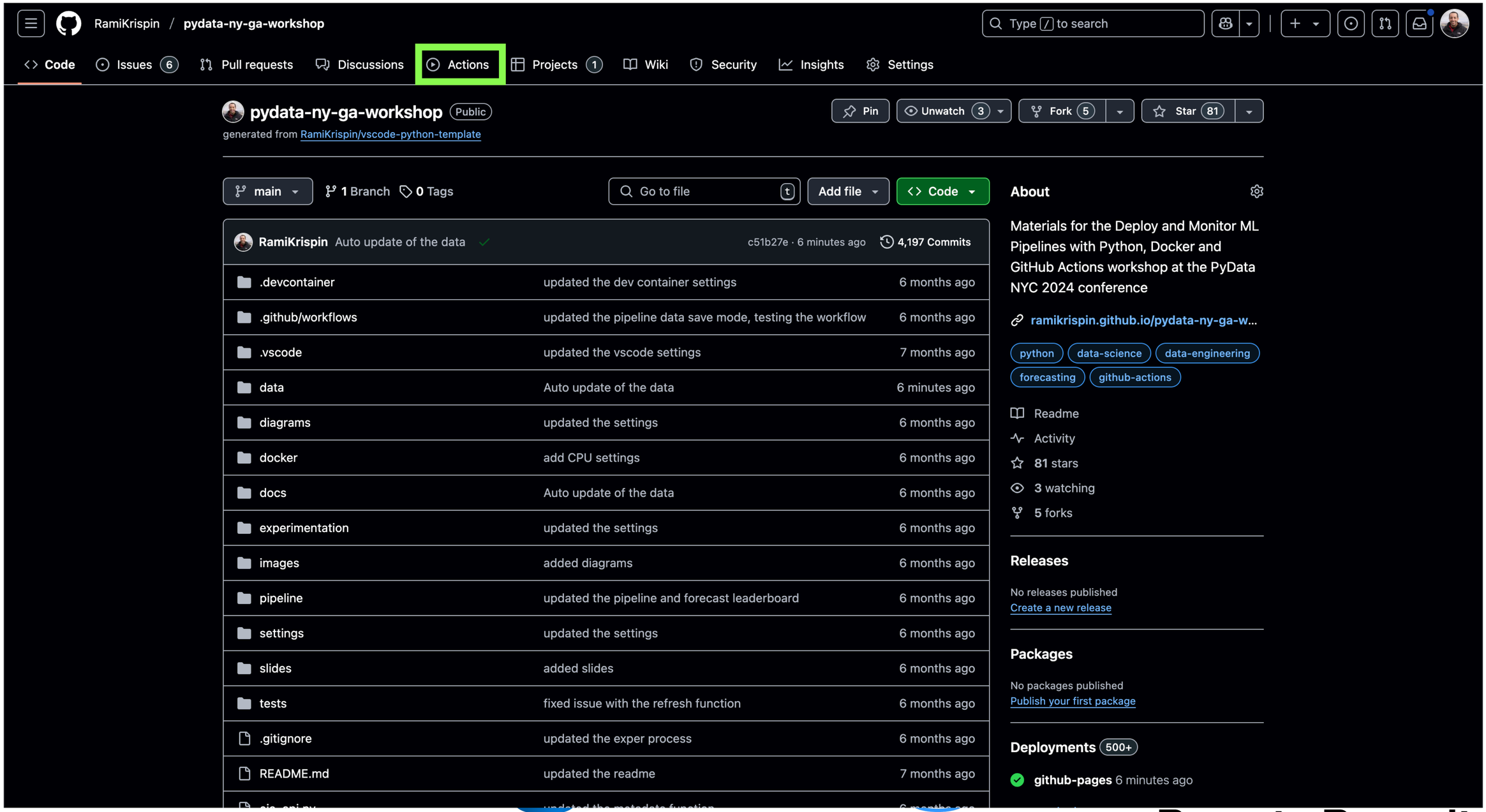
Data & ML Pipeline with GitHub Actions

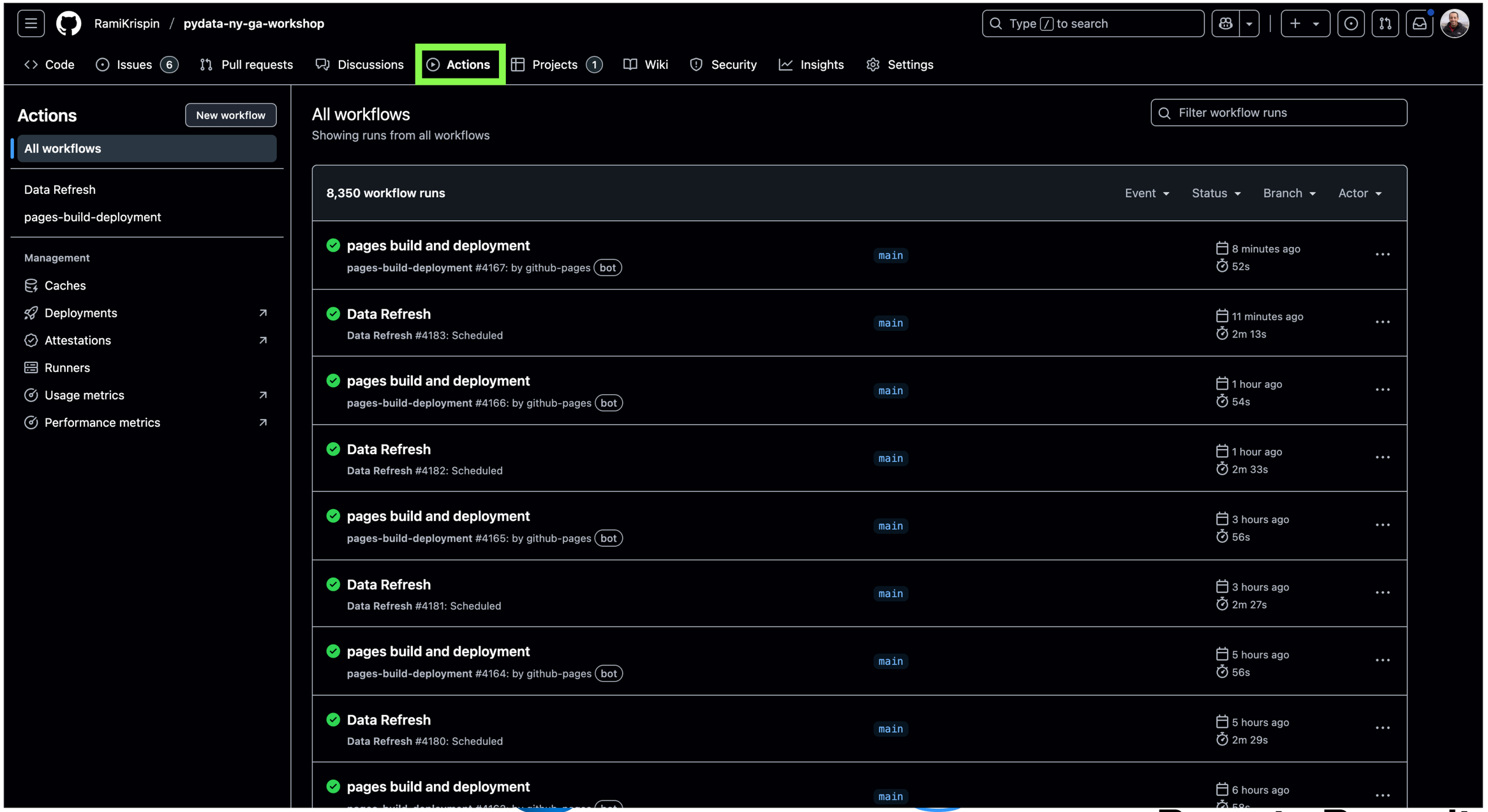
Architecture

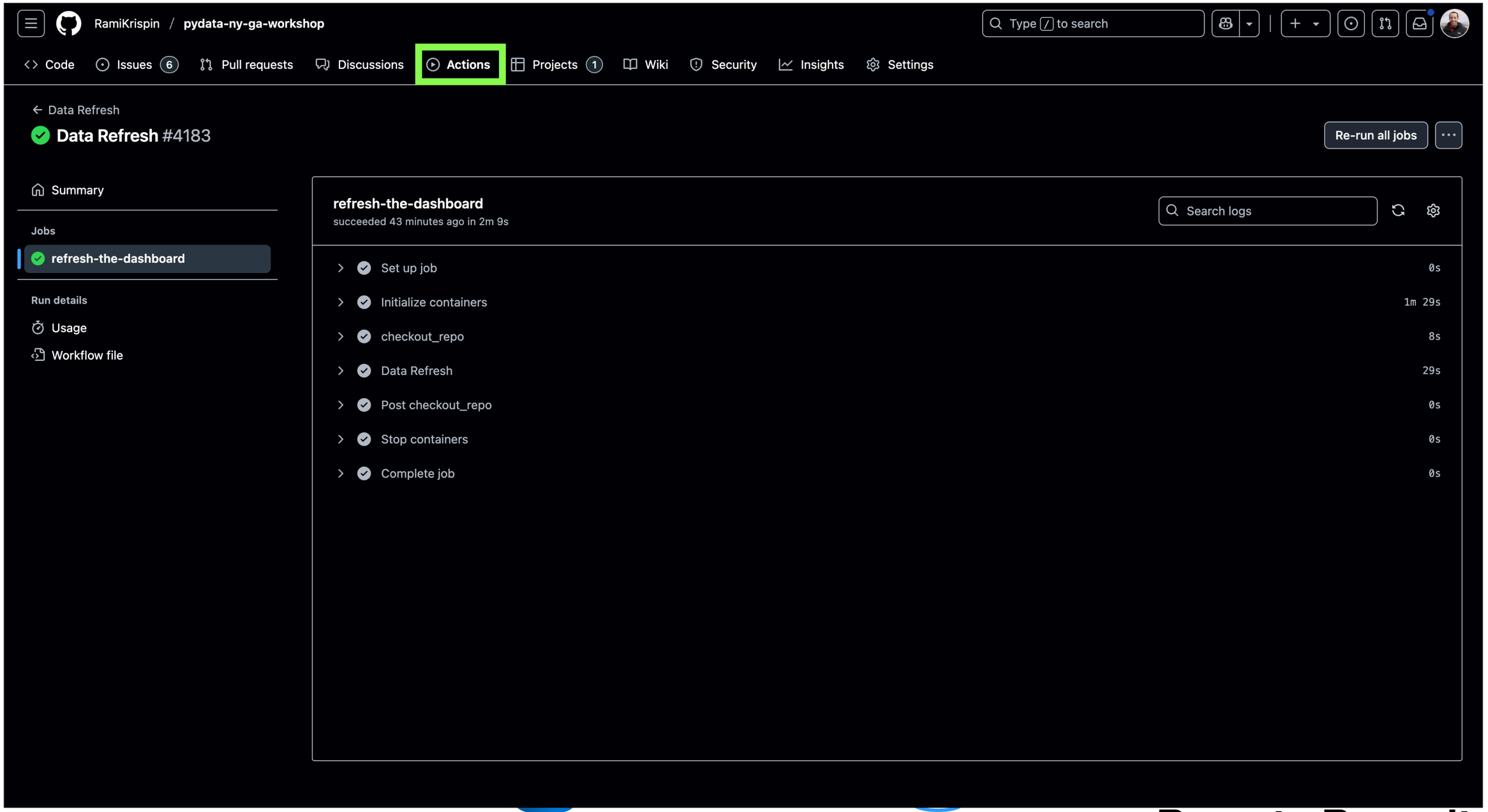


Remote Repository


RamiKrispin / pydata-ny-ga-workshop
Type / to search
Code
Issues 6
Pull requests
Discussions
Actions
Projects 1
Wiki
Security
Insights
Settings
pydata-ny-ga-workshop
Public
generated from RamiKrispin/vscode-python-template
Pin
Unwatch 3
Fork 5
Star 81
main
1 Branch
0 Tags
Go to file
t
Add file
Code
RamiKrispin Auto update of the data
c51b27e · 6 minutes ago
4,197 Commits
.devcontainer updated the dev container settings 6 months ago
.github/workflows updated the pipeline data save mode, testing the workflow 6 months ago
.vscode updated the vscode settings 7 months ago
data Auto update of the data 6 minutes ago
diagrams updated the settings 6 months ago
docker add CPU settings 6 months ago
docs Auto update of the data 6 months ago
experimentation updated the settings 6 months ago
images added diagrams 6 months ago
pipeline updated the pipeline and forecast leaderboard 6 months ago
settings updated the settings 6 months ago
slides added slides 6 months ago
tests fixed issue with the refresh function 6 months ago
.gitignore updated the exper process 6 months ago
README.md updated the readme 7 months ago
About
Materials for the Deploy and Monitor ML Pipelines with Python, Docker and GitHub Actions workshop at the PyData NYC 2024 conference
ramikrispin.github.io/pydata-ny-ga-w...
python
data-science
data-engineering
forecasting
github-actions
Readme
Activity
81 stars
3 watching
5 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Deployments 500+
github-pages 6 minutes ago
Database
S3 Bucket
Remote Repository


RamiKrispin / pydata-ny-ga-workshop
Type / to search
Code
Issues 6
Pull requests
Discussions
Actions
Projects 1
Wiki
Security
Insights
Settings
pydata-ny-ga-workshop Public
generated from RamiKrispin/vscode-python-template
Pin
Unwatch 3
Fork 5
Star 81
main
1 Branch
0 Tags
Go to file
t
Add file
Code
RamiKrispin Auto update of the data
c51b27e · 6 minutes ago
4,197 Commits
.devcontainer updated the dev container settings 6 months ago
.github/workflows updated the pipeline data save mode, testing the workflow 6 months ago
.vscode updated the vscode settings 7 months ago
data Auto update of the data 6 minutes ago
diagrams updated the settings 6 months ago
docker add CPU settings 6 months ago
docs Auto update of the data 6 months ago
experimentation updated the settings 6 months ago
images added diagrams 6 months ago
pipeline updated the pipeline and forecast leaderboard 6 months ago
settings updated the settings 6 months ago
slides added slides 6 months ago
tests fixed issue with the refresh function 6 months ago
.gitignore updated the exper process 6 months ago
README.md updated the readme 7 months ago
About
Materials for the Deploy and Monitor ML Pipelines with Python, Docker and GitHub Actions workshop at the PyData NYC 2024 conference
ramikrispin.github.io/pydata-ny-ga-w...
python
data-science
data-engineering
forecasting
github-actions
Readme
Activity
81 stars
3 watching
5 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Deployments 500+
github-pages 6 minutes ago
Database
S3 Bucket
Remote Repository


RamiKrispin / pydata-ny-ga-workshop
Type / to search
Code
Issues 6
Pull requests
Discussions
Actions
Projects 1
Wiki
Security
Insights
Settings
Actions
New workflow
All workflows
Data Refresh
pages-build-deployment
Management
Caches
Deployments
Attestations
Runners
Usage metrics
Performance metrics
All workflows
Showing runs from all workflows
Filter workflow runs
8,350 workflow runs
Event
Status
Branch
Actor
pages build and deployment
pages-build-deployment #4167: by github-pages bot
main
8 minutes ago
52s
Data Refresh
Data Refresh #4183: Scheduled
main
11 minutes ago
2m 13s
pages build and deployment
pages-build-deployment #4166: by github-pages bot
main
1 hour ago
54s
Data Refresh
Data Refresh #4182: Scheduled
main
1 hour ago
2m 33s
pages build and deployment
pages-build-deployment #4165: by github-pages bot
main
3 hours ago
56s
Data Refresh
Data Refresh #4181: Scheduled
main
3 hours ago
2m 27s
pages build and deployment
pages-build-deployment #4164: by github-pages bot
main
5 hours ago
56s
Data Refresh
Data Refresh #4180: Scheduled
main
5 hours ago
2m 29s
pages build and deployment
main
6 hours ago
Database
S3 Bucket
Remote Repository



Database

S3 Bucket

Remote Repository

Data & ML Pipeline with GitHub Actions

Summary

- CI/CD platform
- Out-of-the-box
- Automate workflows
- Deployment with Docker
- Both free and enterprise version

Questions?

Thank You!