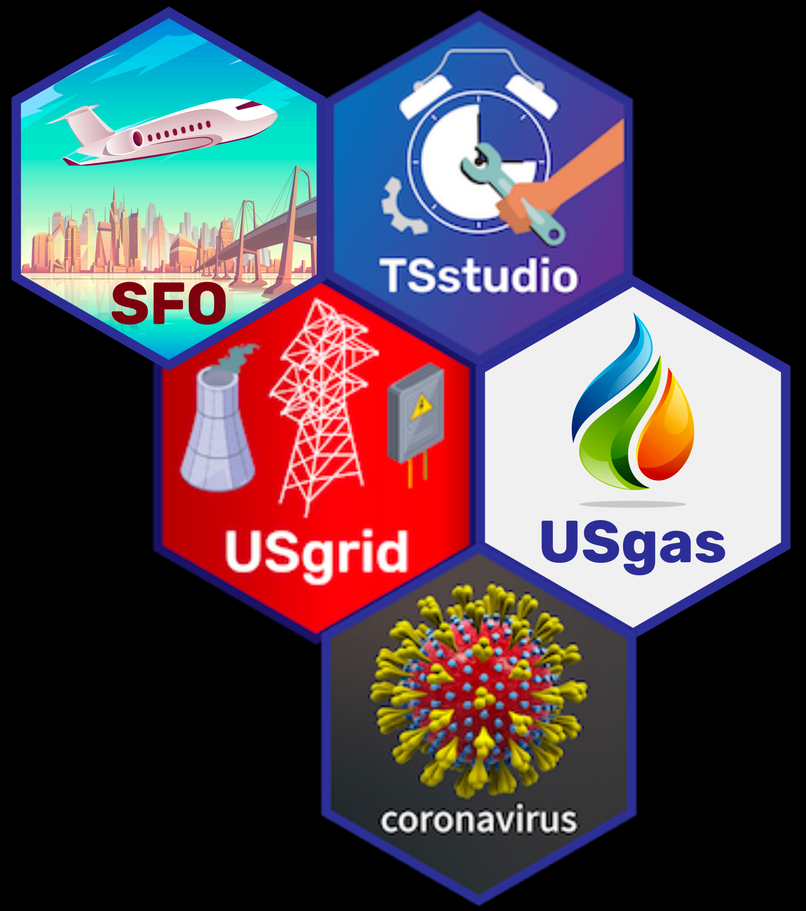
Deploy and Monitor ML Pipelines with Python, Docker, and GitHub Actions

PyData NYC 2024 Conference

Rami Krispin, November 6th, 2024

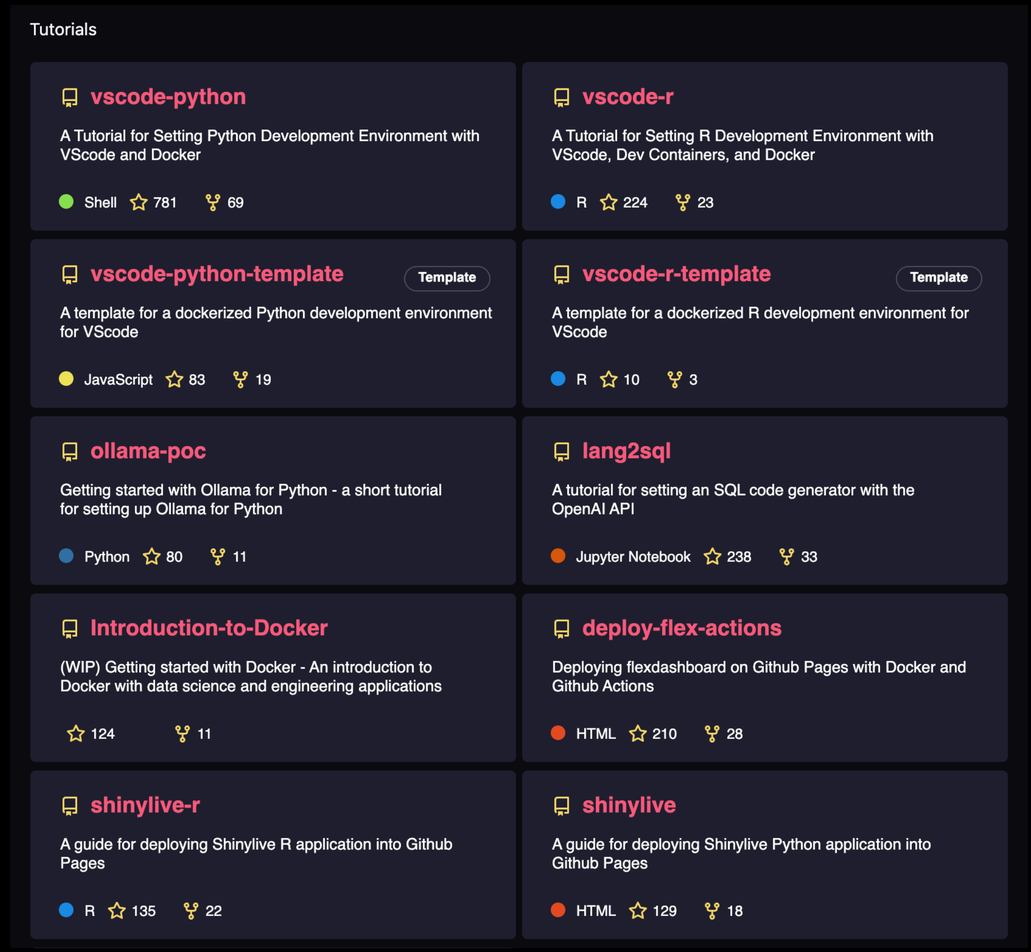
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 🐳



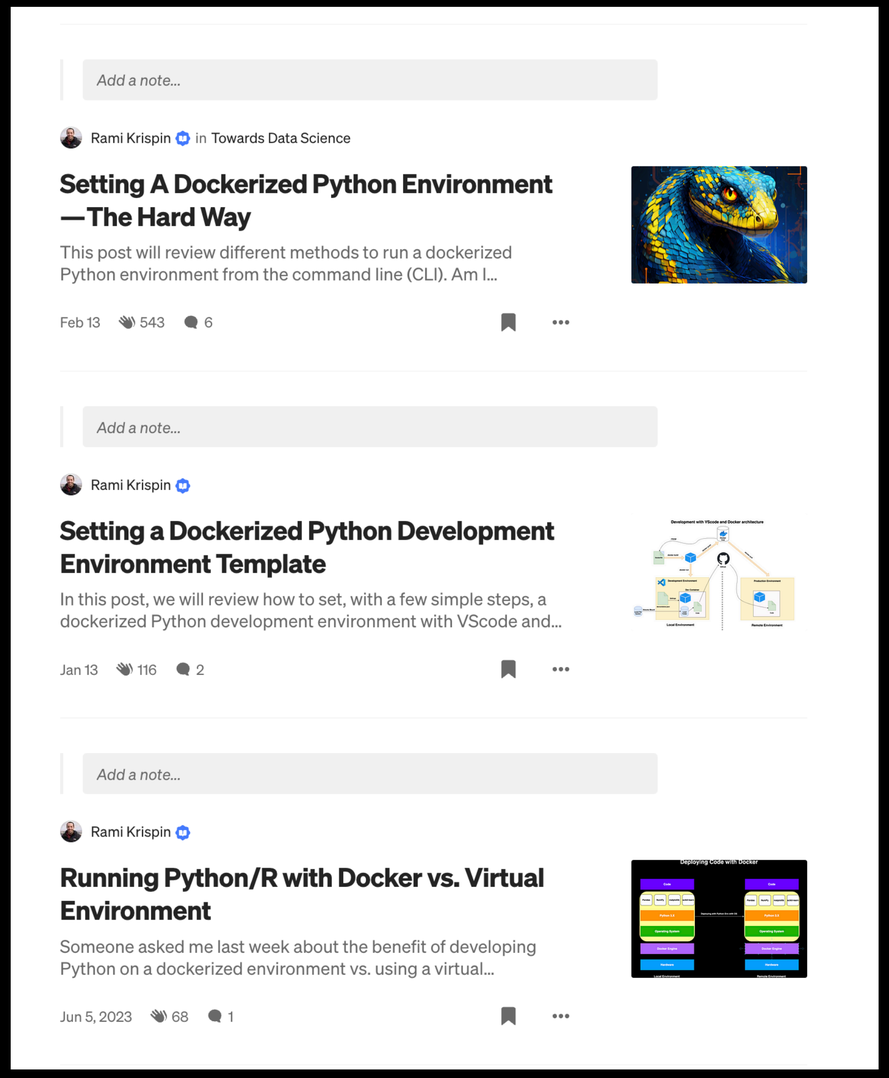
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 🐳



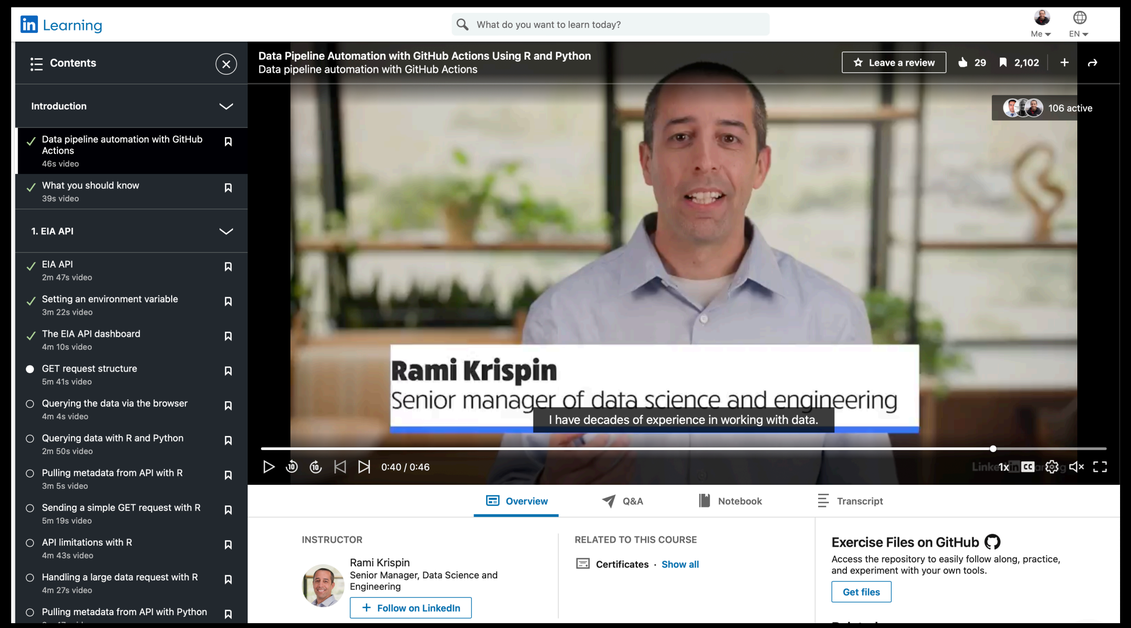
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 🐳



About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author
- Docker Captain 🐳



About Me

- Senior Manager
- **Forecasting**
- **MLOps**
- **Open Source**
- Author
- Docker Captain 🐋

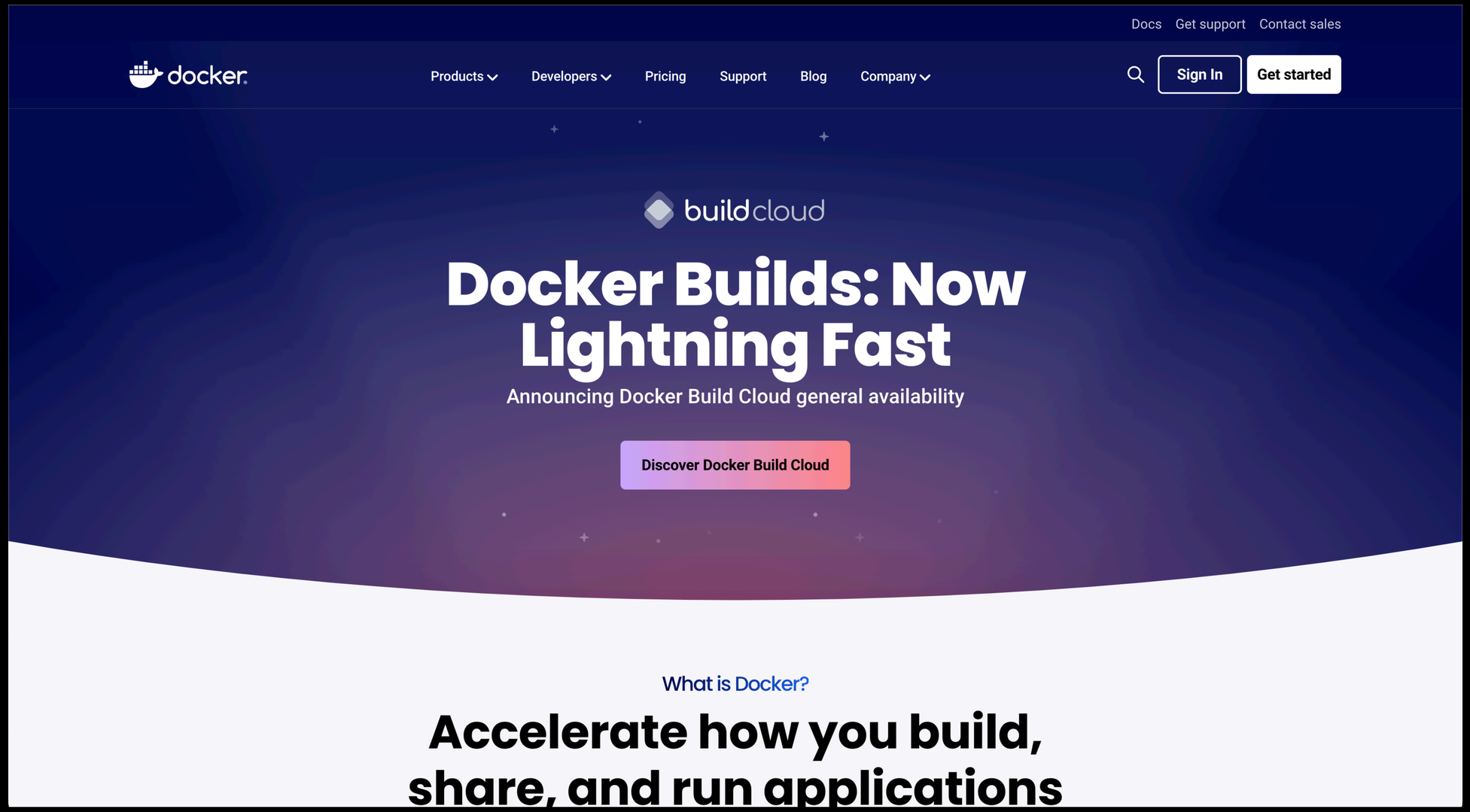
https://github.com/RamiKrispin/pydata-ny-ga-workshop/

Poll

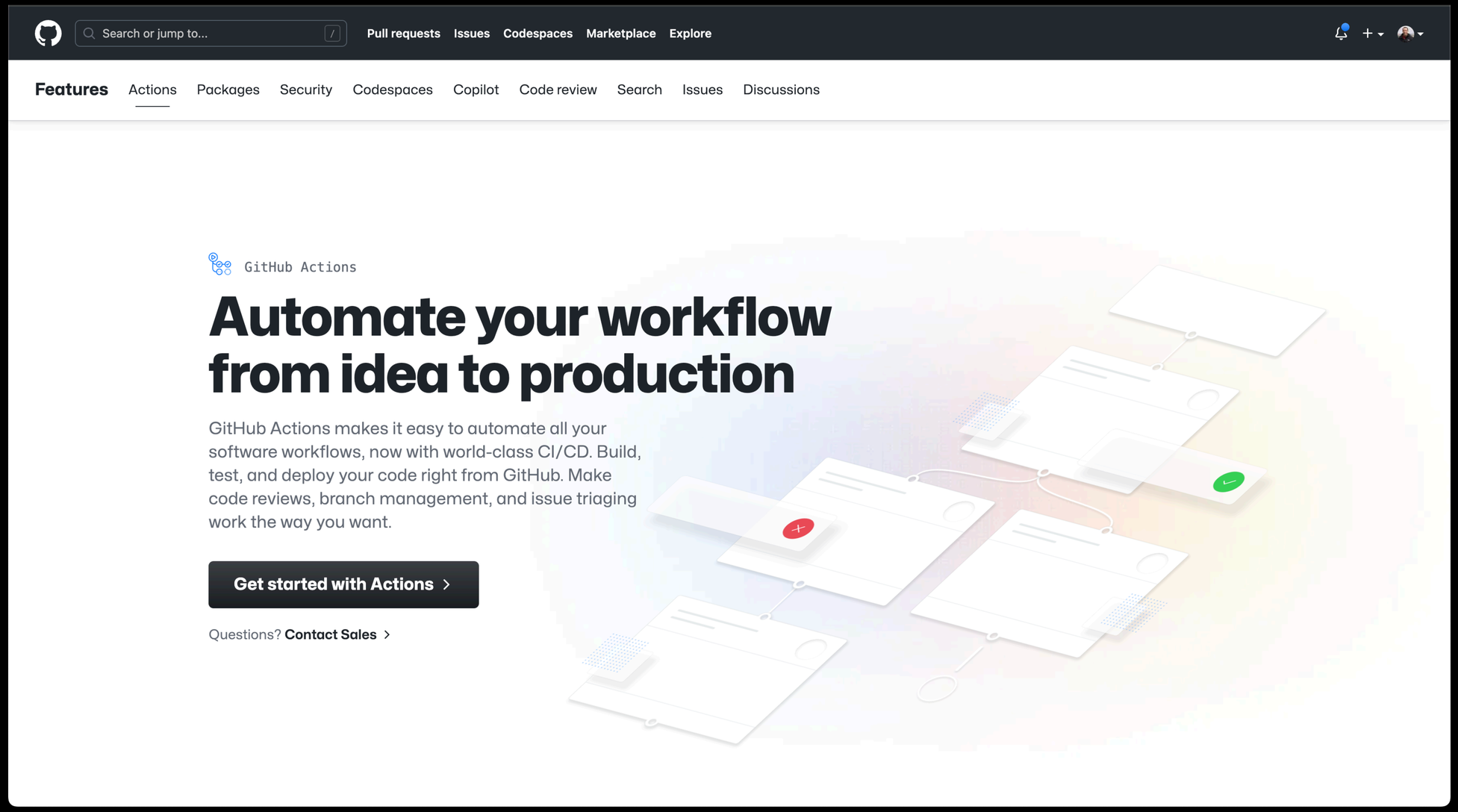
Are You Familiar with?

- Docker
- GitHub Actions
- Quarto
- Forecasting

Docker


Docs
Get support
Contact sales
docker
Products
Developers
Pricing
Support
Blog
Company
Sign In
Get started
buildcloud
Docker Builds: Now Lightning Fast
Announcing Docker Build Cloud general availability
Discover Docker Build Cloud
What is Docker?
Accelerate how you build, share, and run applications

GitHub Actions

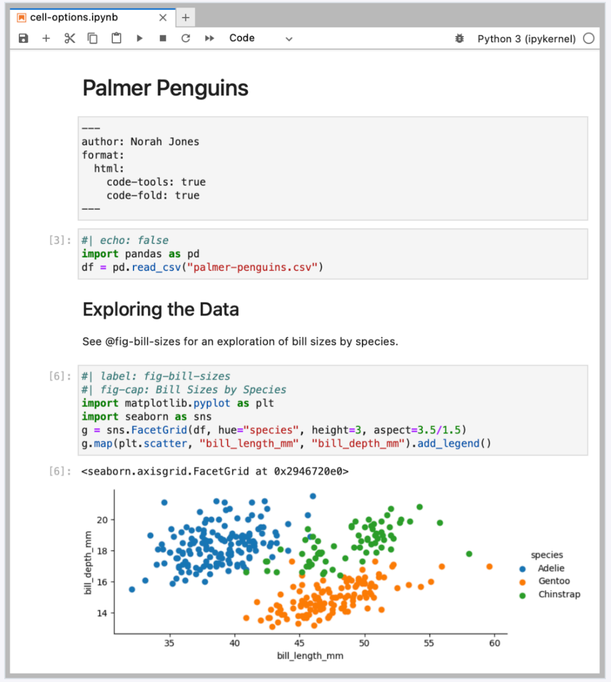


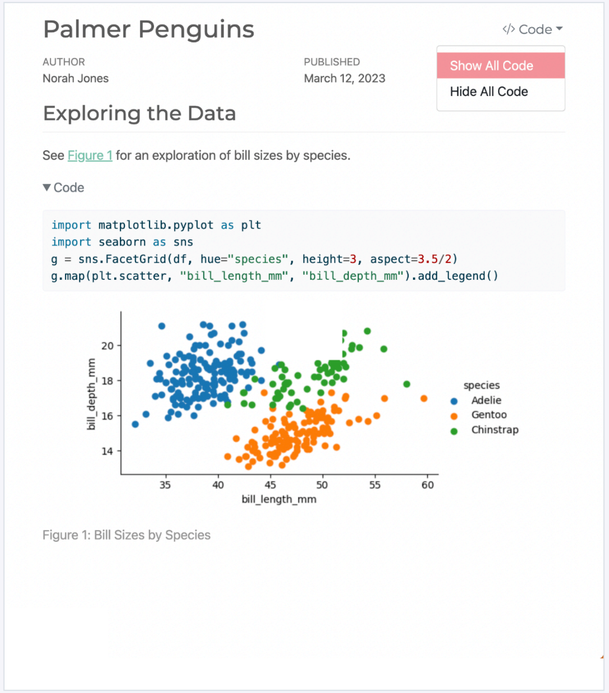
Quarto

Hello, Quarto

Python R Julia Observable

Combine Jupyter notebooks with flexible options to produce production quality output in a wide variety of formats. Author using traditional notebook UIs or with a plain text markdown representation of notebooks.





Dynamic Documents

Generate dynamic output using Python, R, Julia,

Beautiful Publications

Publish high-quality articles, reports,

Scientific Markdown

Pandoc markdown has excellent support for

Nixtla mlforecast

NIXTLA

Search or ask...

Hyperparameter optimization
Transforming exogenous features
Cross validation
Probabilistic forecasting
Target transformations
Using scikit-learn pipelines
Analyzing the trained models
MLflow
Custom training
Training with numpy arrays
One model per step
Custom date features
Predict callbacks
Predicting a subset of ids
Transfer Learning

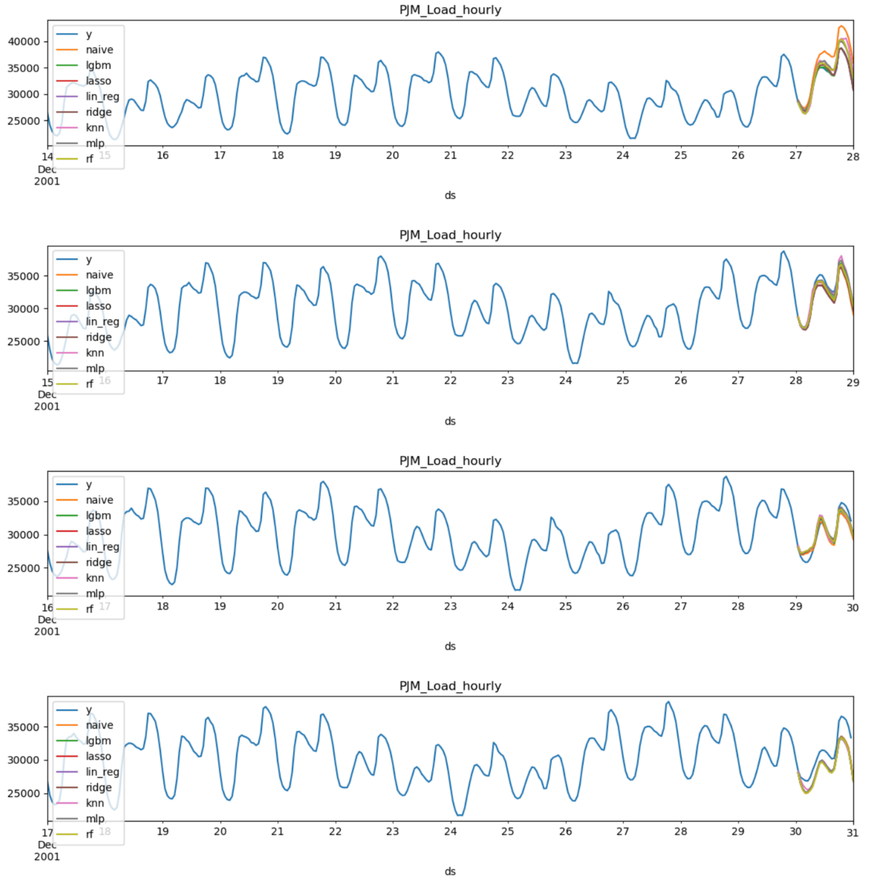
Tutorials

Electricity Load Forecast
Detect Demand Peaks
Prediction intervals

API Reference

Local
Distributed

```
plot_cv(df_train, crossvalidation_df, 'PJM_Load_hourly', 'load_forecasting__pr
```



On this page

Introduction
Libraries
Forecast using Multiple Seasonalities
Electricity Load Data
Analizing Seasonalities
Model Selection with Cross-Validation
Test Evaluation
Comparison with Prophet

Ask AI

Alternatives Tools

Agenda

Motivation

Workflow

Design

Demo

Motivation

February 2020

Chinese stocks plunged 8% as coronavirus fears took hold. It's the worst day in years

By Laura He, CNN Business

Updated 9:15 AM EST, Mon February 3, 2020

Source:
CNN

10:32 p.m. ET, February 2, 2020

The best photos from the game

Kyle Terada/USA Today Sports

Source:
CNN

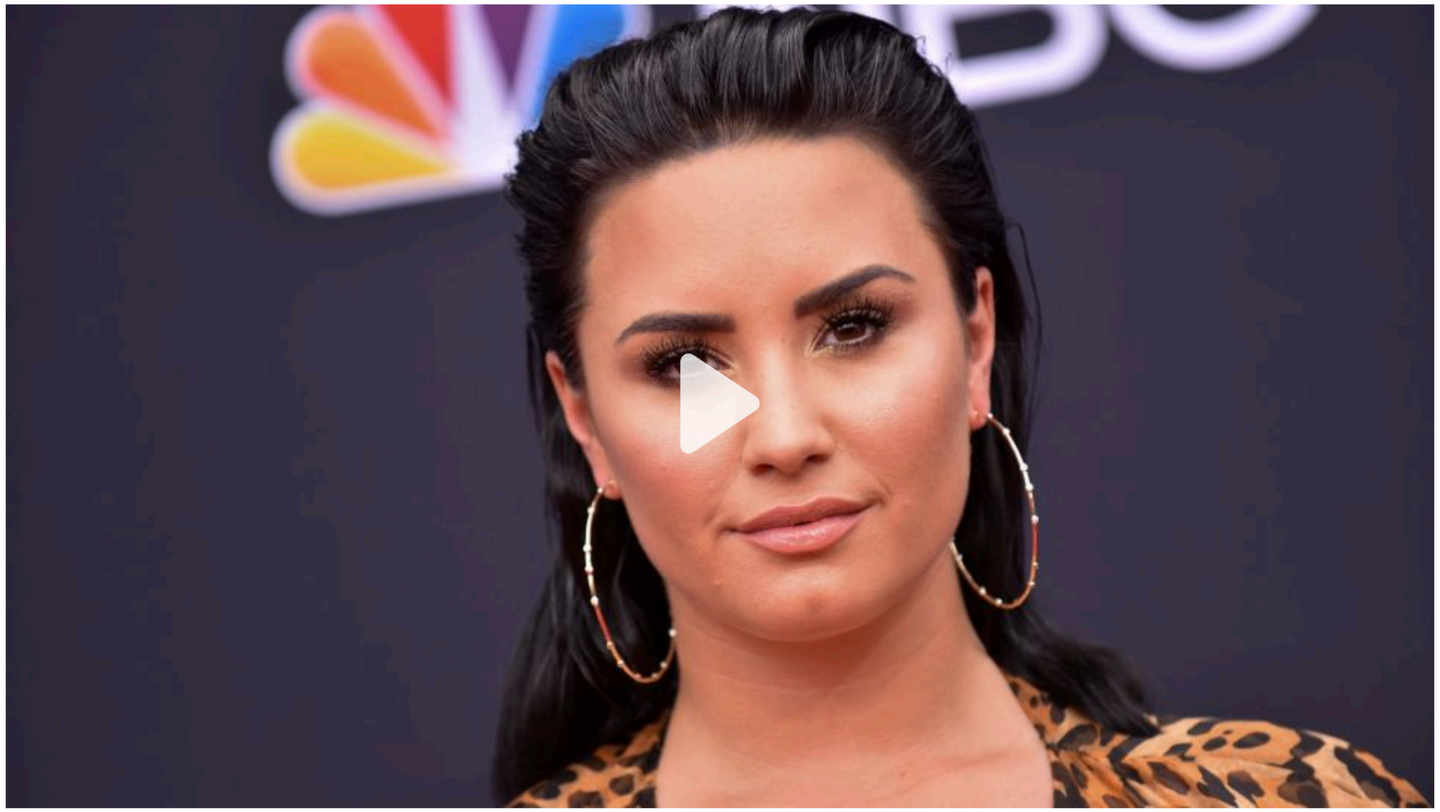
Jennifer Lopez's epic Super Bowl flag coat was custom Versace

By Sandra Gonzalez, CNN
Updated 6:58 AM EST, Mon February 3, 2020

Source:
CNN

Demi Lovato rocks the National Anthem at Super Bowl LIV

By Lisa Respers France, CNN
Updated 11:27 PM EST, Sun February 2, 2020



Source:
CNN

Trump and Pelosi haven’t spoken in months

By Manu Raju, Senior Congressional Correspondent
Updated 8:49 AM EST, Mon February 3, 2020

Source:
CNN

‘No reason for Americans to panic’: White House seeks to calm fears over coronavirus

By Chandelis Duster, CNN

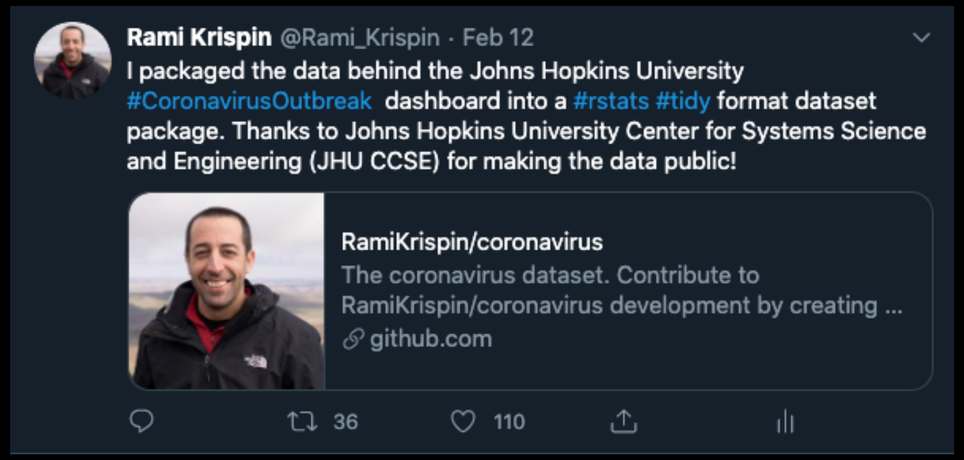
Updated 6:21 PM EST, Sun February 2, 2020

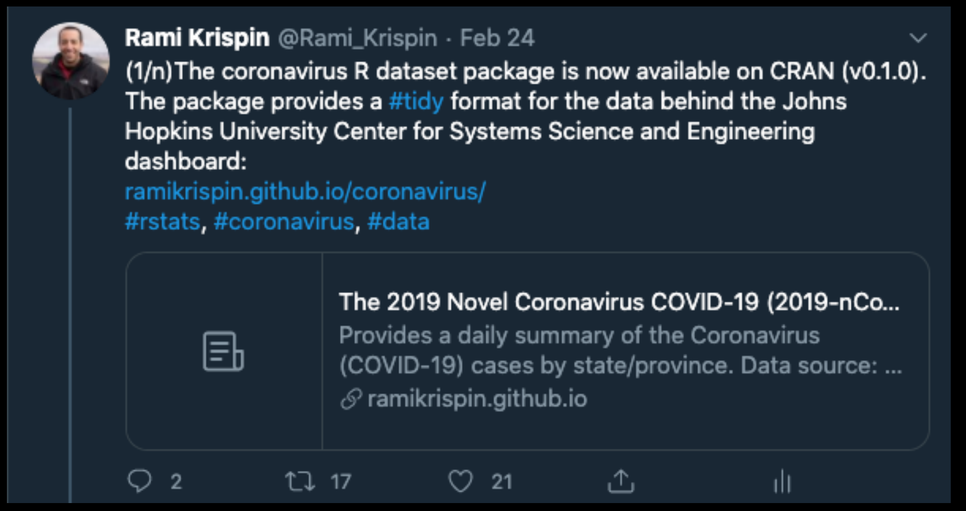
Source:
CNN


EVERYBODY PANIC!


TYRE CENTRE


Rami Krispin @Rami_Krispin · Feb 8
Does anyone aware of a public dataset of the #coronavirus by case over time (e.g., time, city, country, lon, lat, etc...)?
2
1
1


Rami Krispin @Rami_Krispin · Feb 12
I packaged the data behind the Johns Hopkins University #CoronavirusOutbreak dashboard into a #rstats #tidy format dataset package. Thanks to Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) for making the data public!
RamiKrispin/coronavirus
The coronavirus dataset. Contribute to RamiKrispin/coronavirus development by creating ...
github.com
36
110

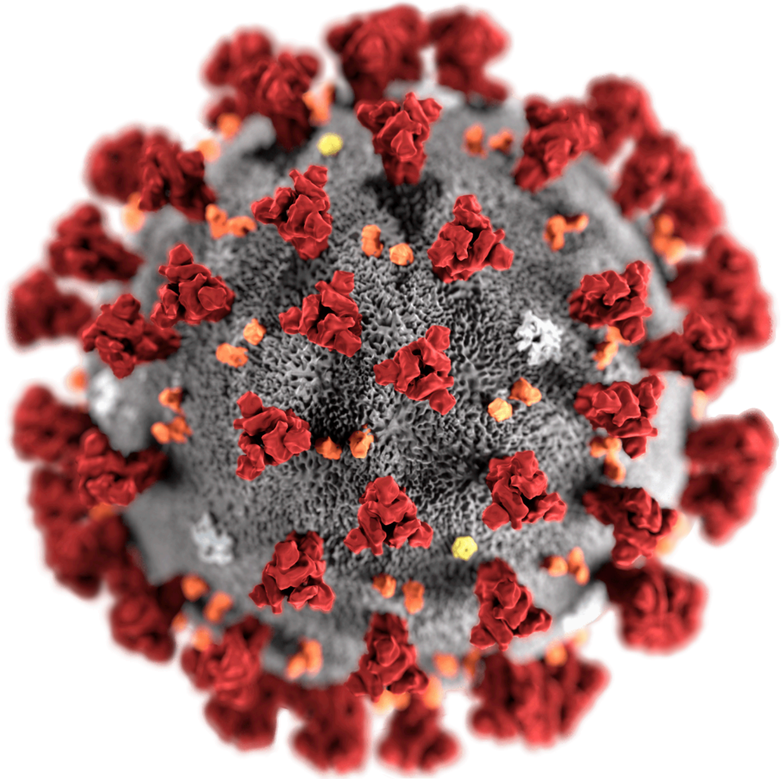

Rami Krispin @Rami_Krispin · Feb 24
(1/n)The coronavirus R dataset package is now available on CRAN (v0.1.0). The package provides a #tidy format for the data behind the Johns Hopkins University Center for Systems Science and Engineering dashboard:
ramikrispin.github.io/coronavirus/
#rstats, #coronavirus, #data
The 2019 Novel Coronavirus COVID-19 (2019-nCo...
Provides a daily summary of the Coronavirus (COVID-19) cases by state/province. Data source: ...
ramikrispin.github.io
2
17
21

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

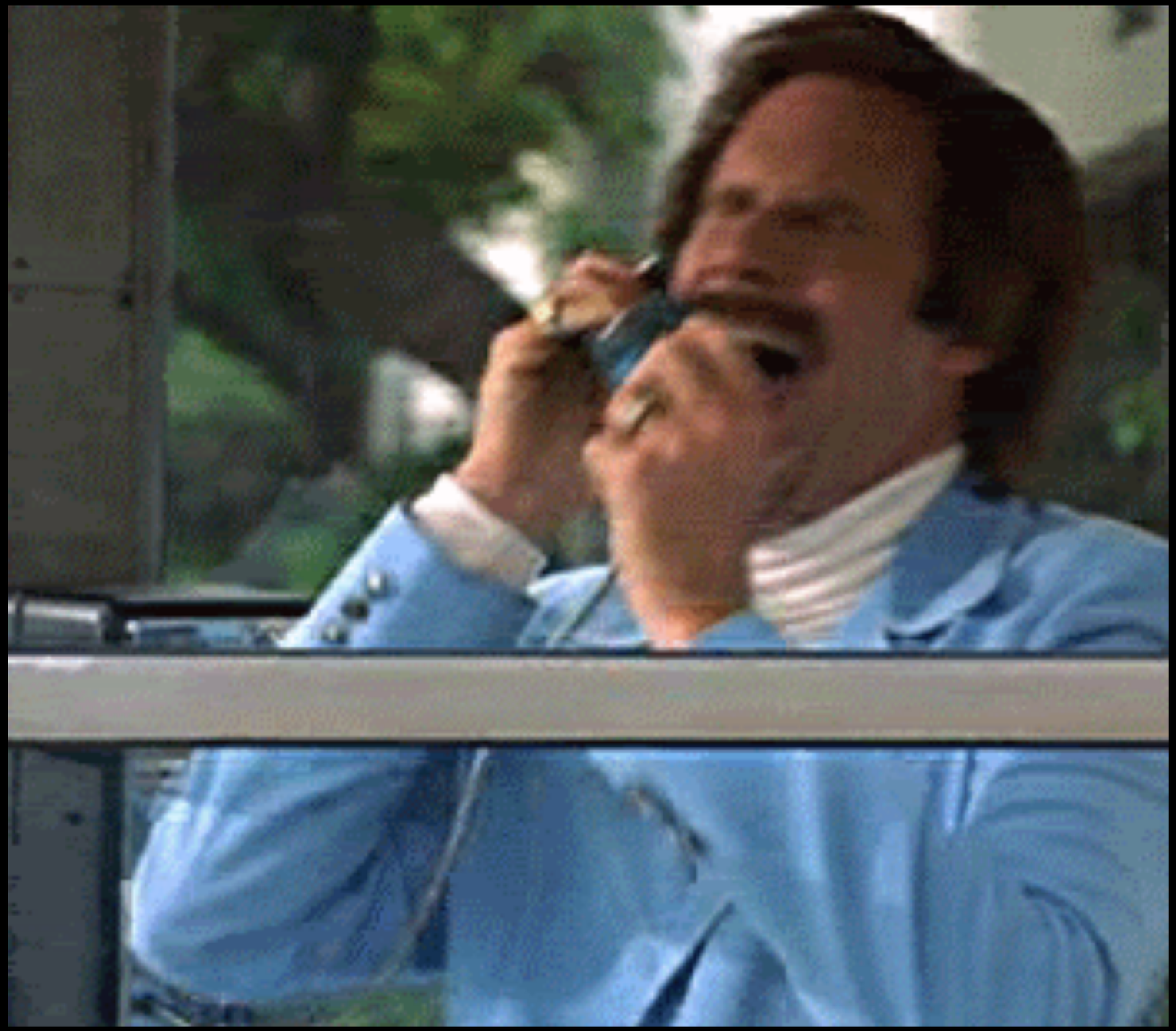
lifecycle stable

License MIT

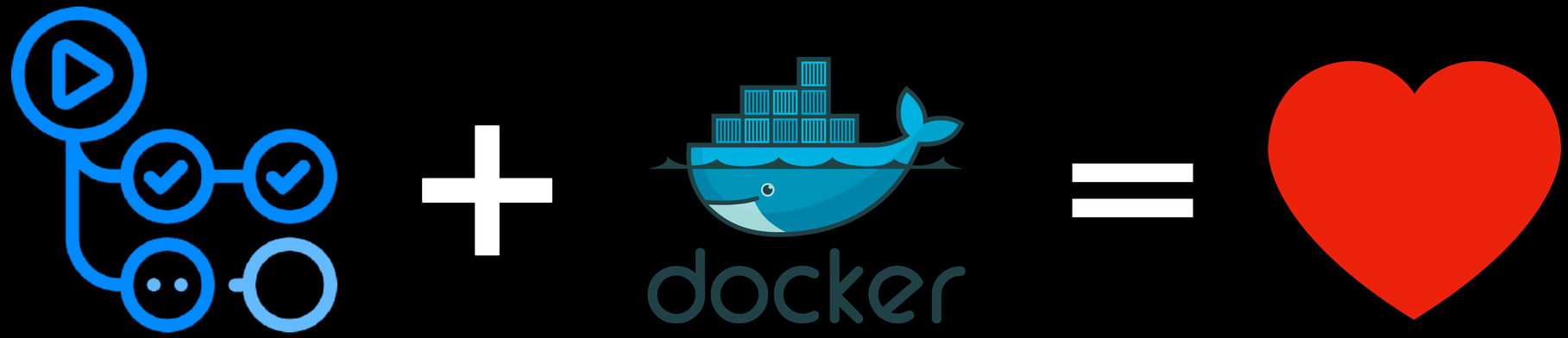
last commit march

downloads 74K

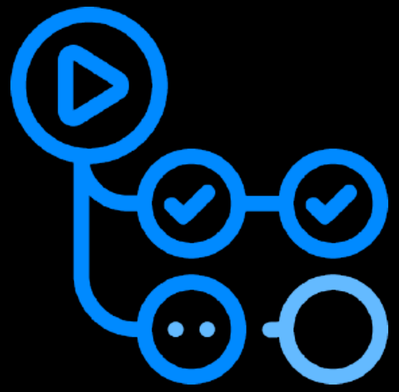
- **Connecting through Power BI** — 1
  #14 by jimmyd7377 was closed on Mar 20, 2020
- **Error when devtools::install_github("RamiKrispin/coronavirus")** — 1
  #13 by Danielchui was closed on Mar 20, 2020
- **Error in Hubei data for 3/11/2020** — 2
  #12 by tcarleton was closed on Mar 13, 2020
- **Country 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



Actions + Docker

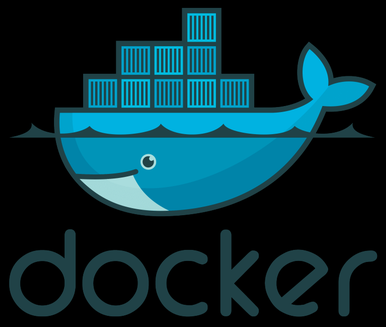


Actions + Docker



GitHub Actions

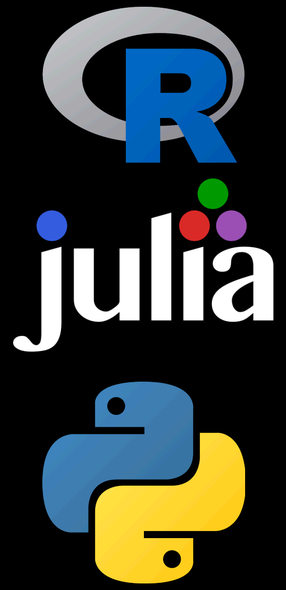
- Scheduler
- CI/CD

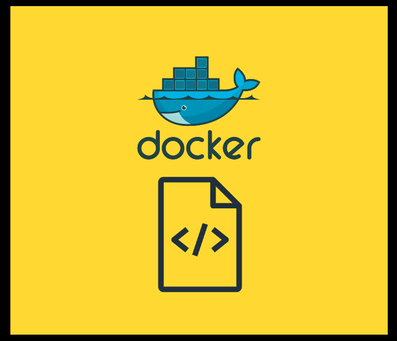


Docker

- Container
- CI/CD

Docker In A Nutshell

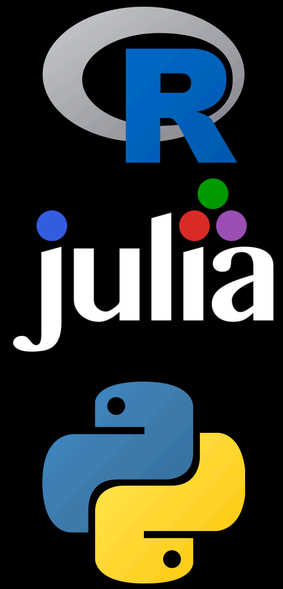


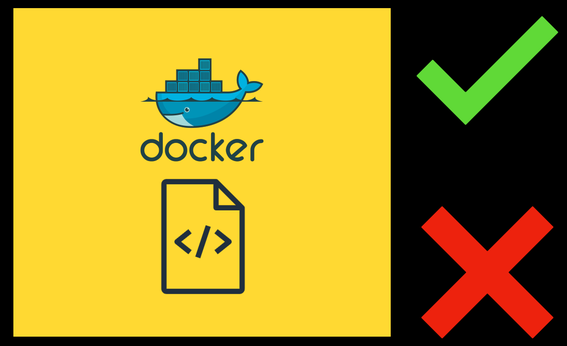


Local Env

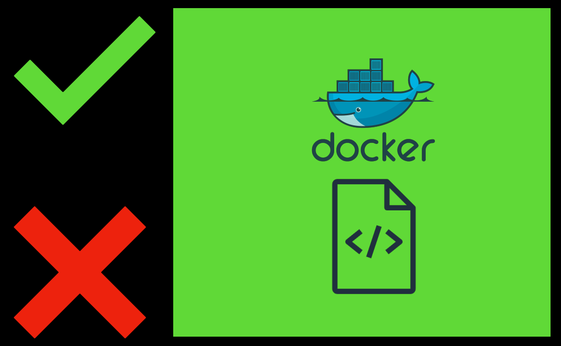
Remote Env

Docker In A Nutshell





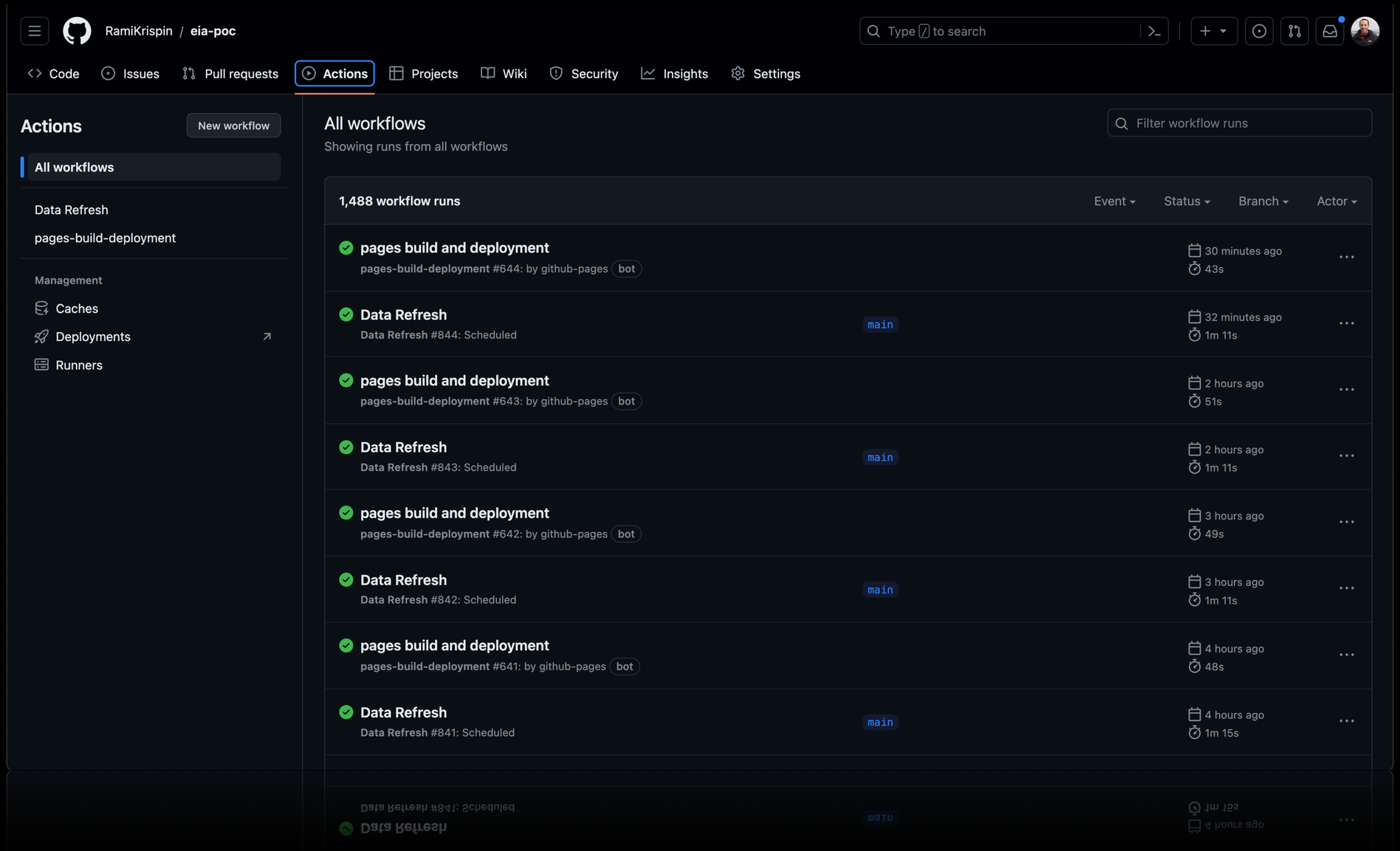
Local Env



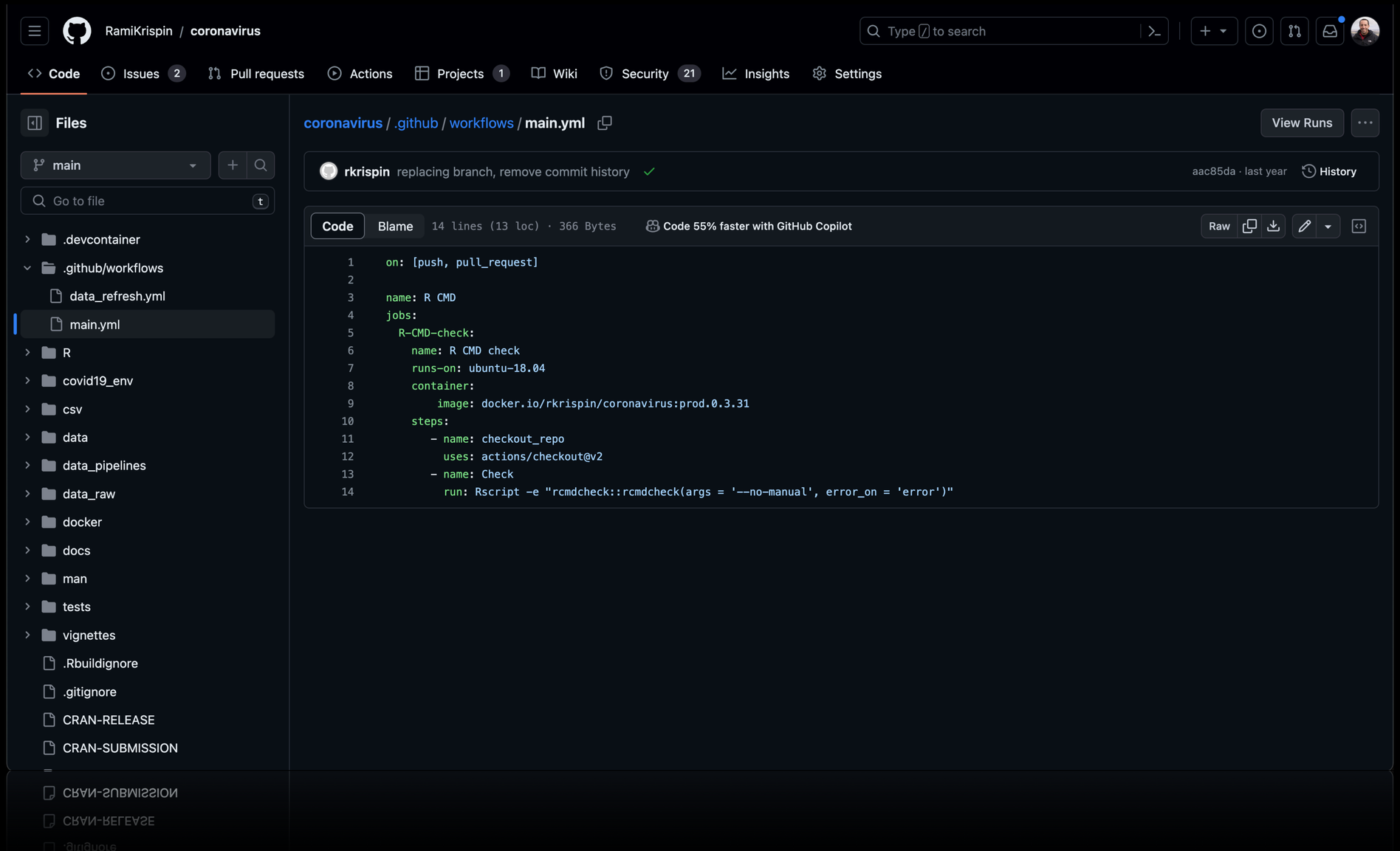
Remote Env

What Can You Build?

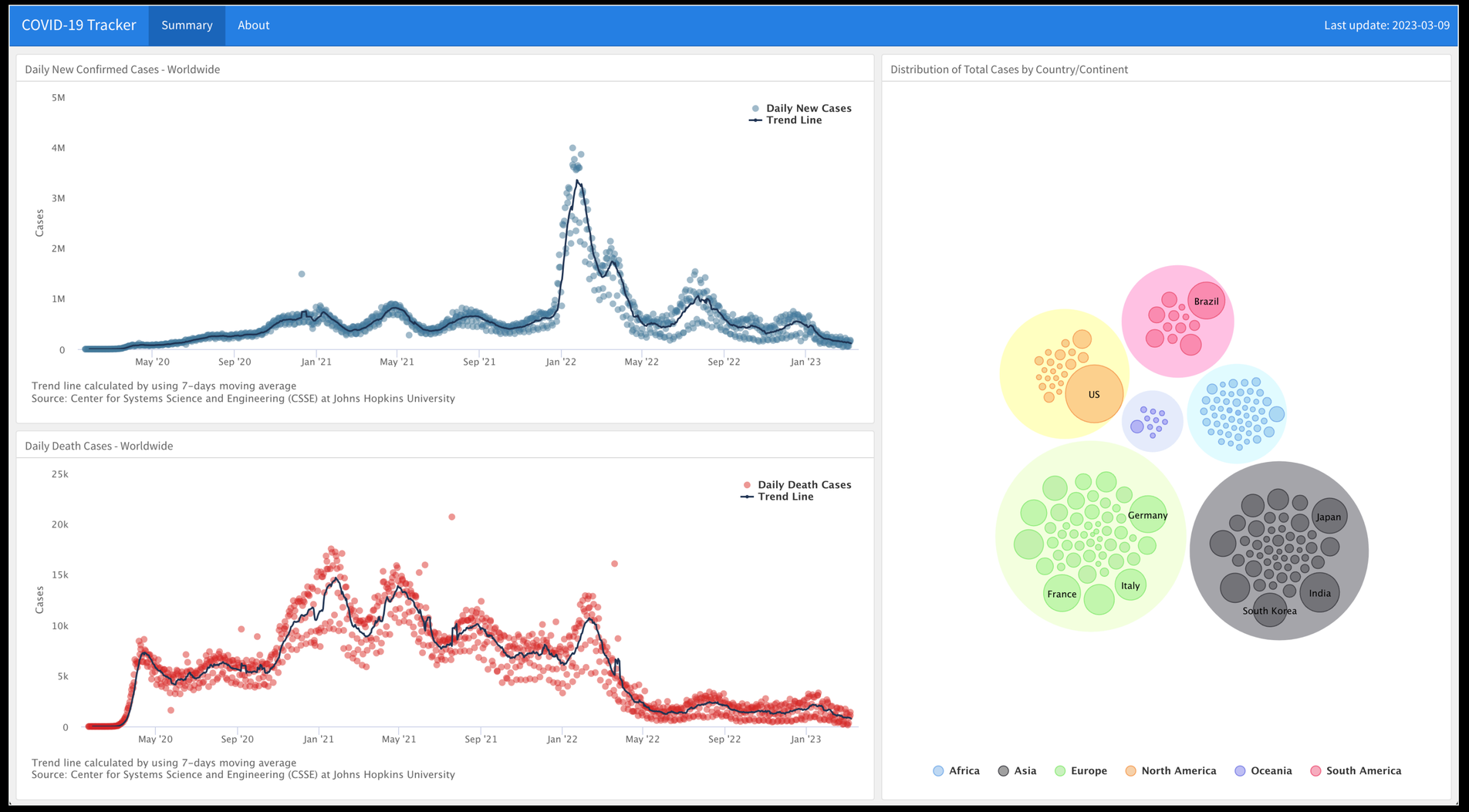
CI/CD & Automation



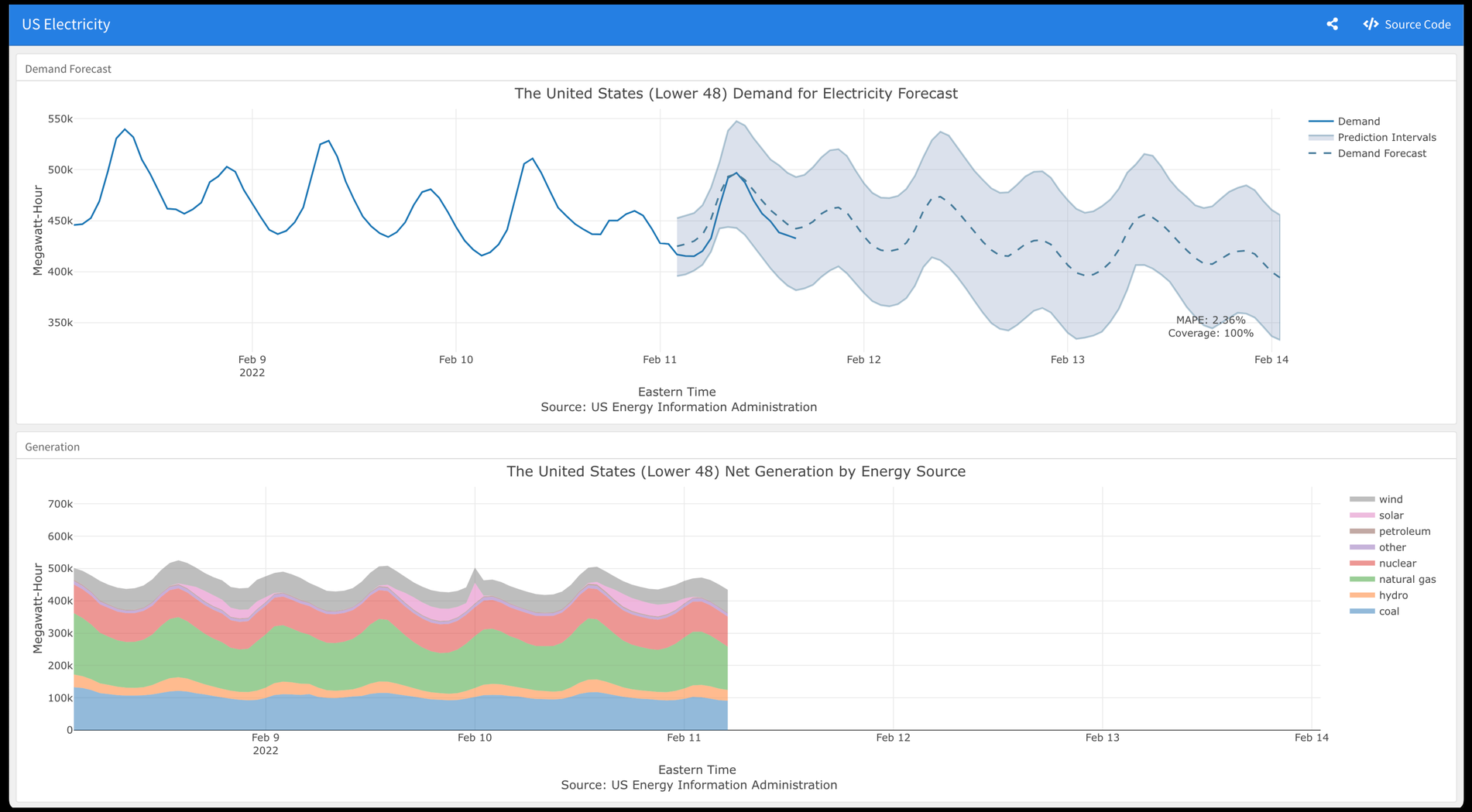
Code Development and Testing



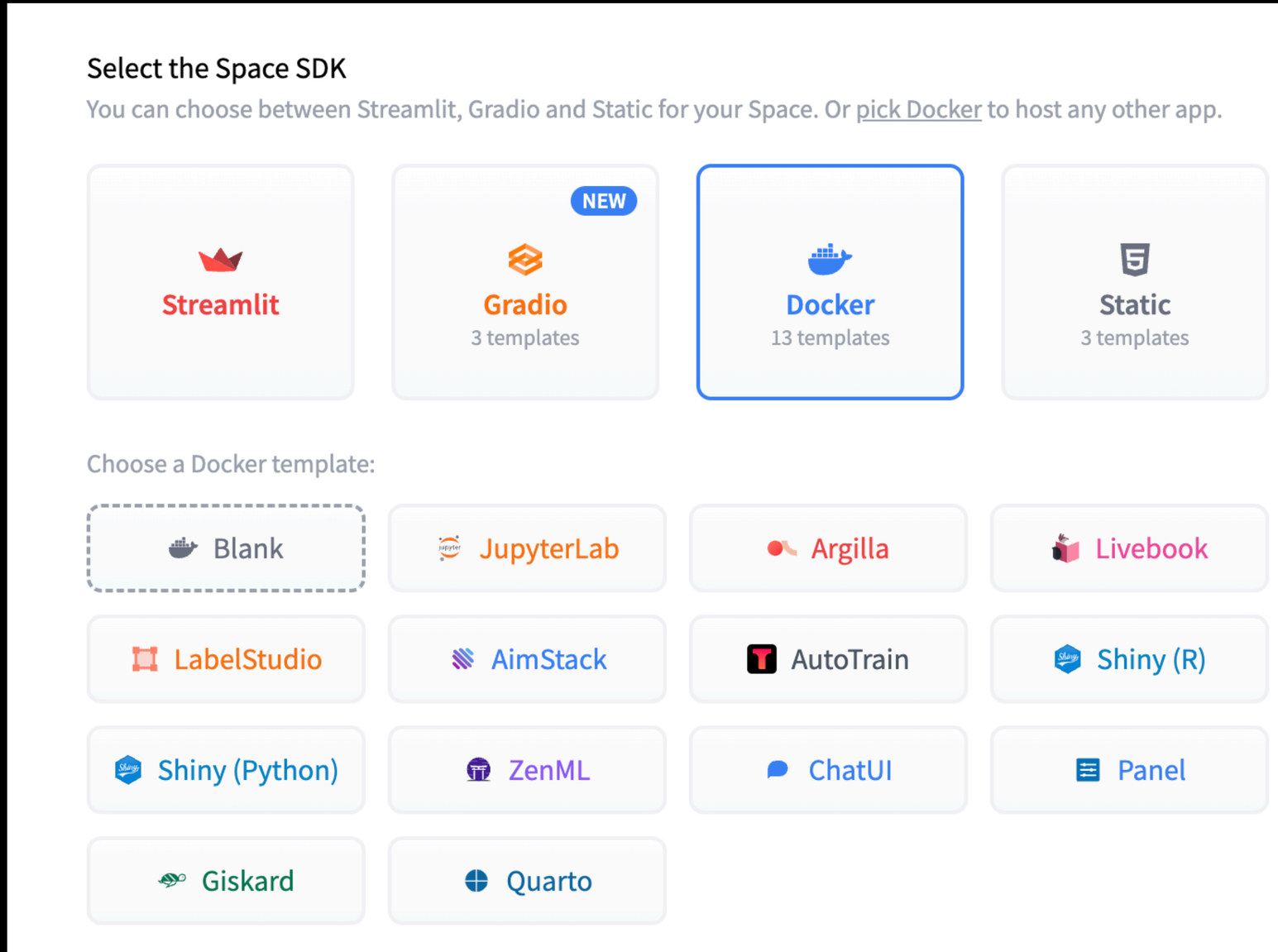
Data Automation



ML Applications



Hugging Face 🤗



Open and Free To Use

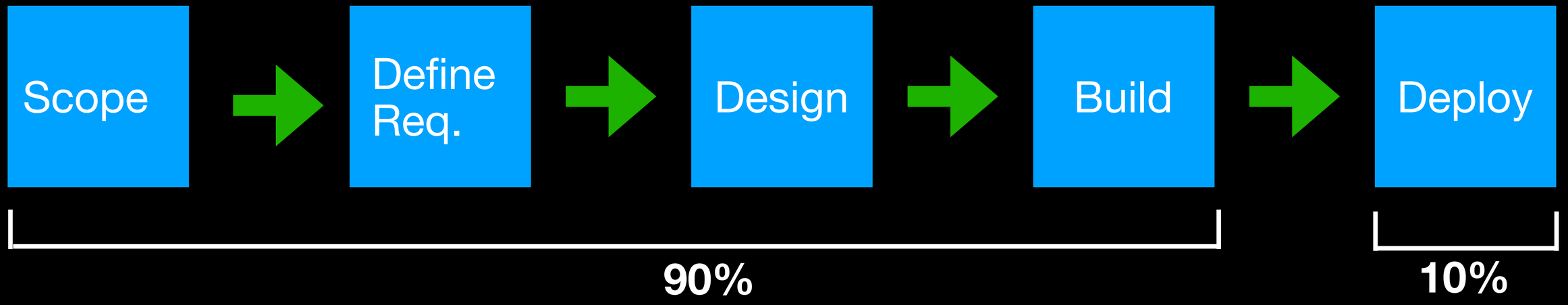
Limitations

Limitations

- Not secure
- Compute power
- Availability
- Data storage

Deploy and Monitor ML and Data Pipelines with GitHub Actions

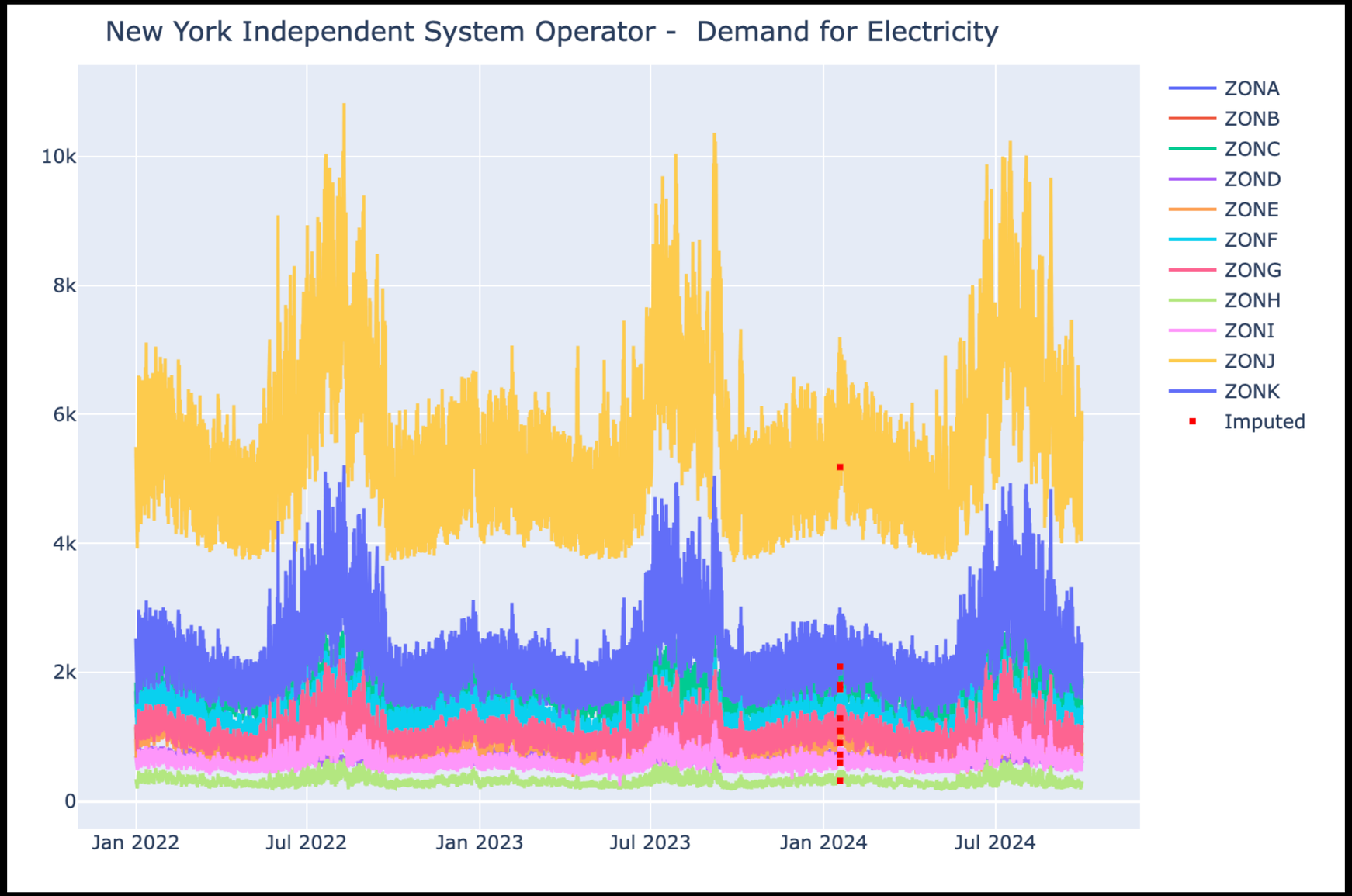
Building Pipeline - Workflow



Scope

- Forecast the hourly demand for electricity in New York by sub-region
- This includes the following 11 providers:
- Refresh the data daily
- Update the forecast every 24 hours

Scope

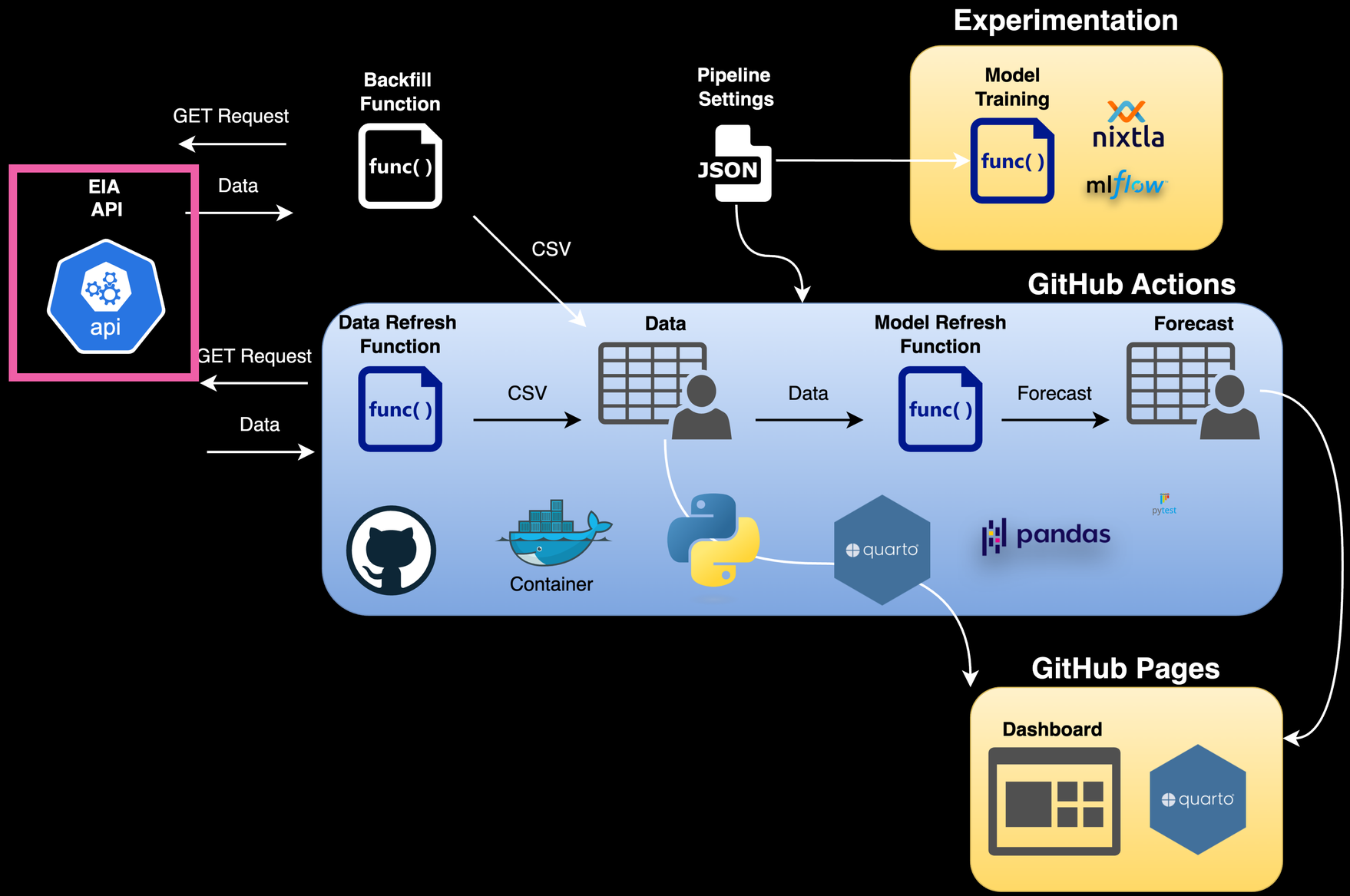

New York Independent System Operator - Demand for Electricity
10k
8k
6k
4k
2k
0
Jan 2022
Jul 2022
Jan 2023
Jul 2023
Jan 2024
Jul 2024
ZONA
ZONB
ZONC
ZOND
ZONE
ZONF
ZONG
ZONH
ZONI
ZONJ
ZONK
Imputed

Requirements

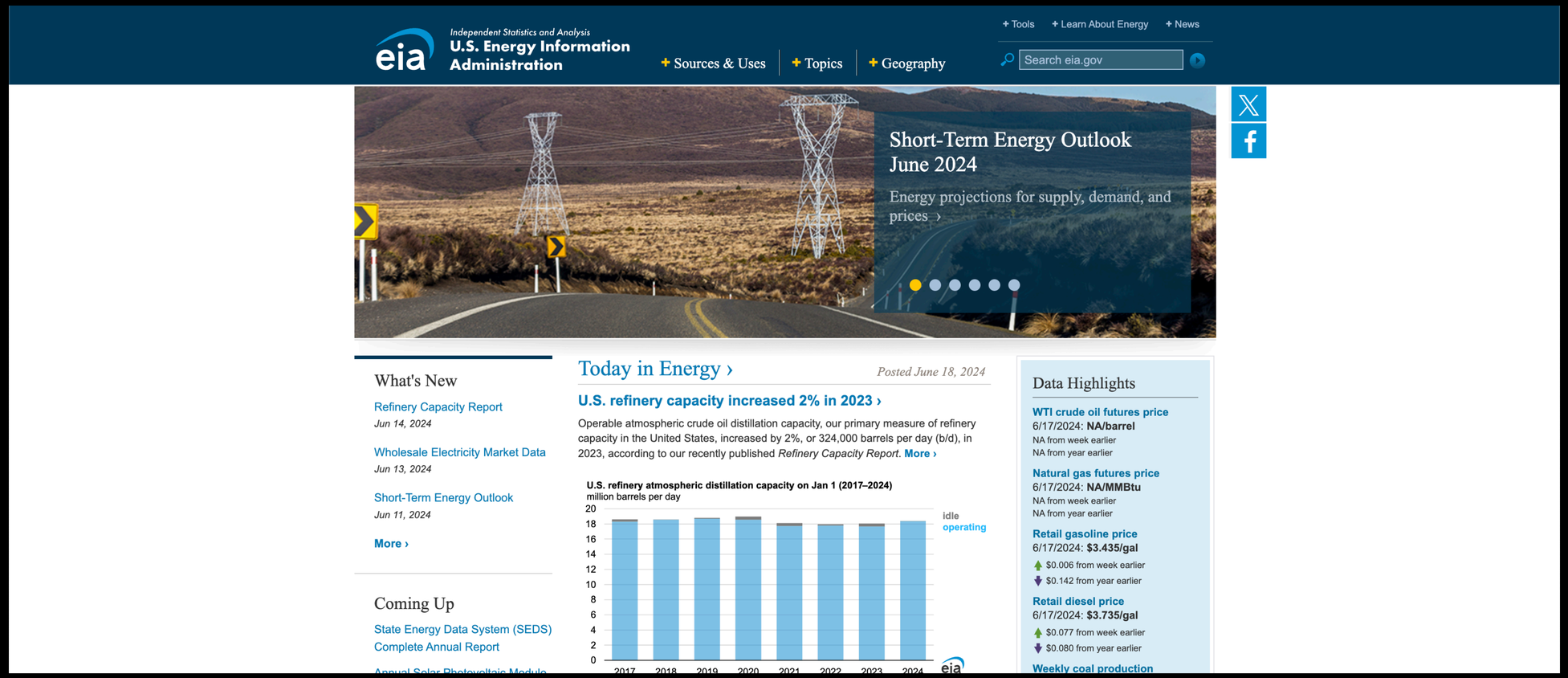
- Environment - Docker image
- Orchestration - GitHub Actions
- Data pipeline - keep the data up-to-date
- ML pipeline - refresh the forecast
- Dashboard - GitHub Pages

Design

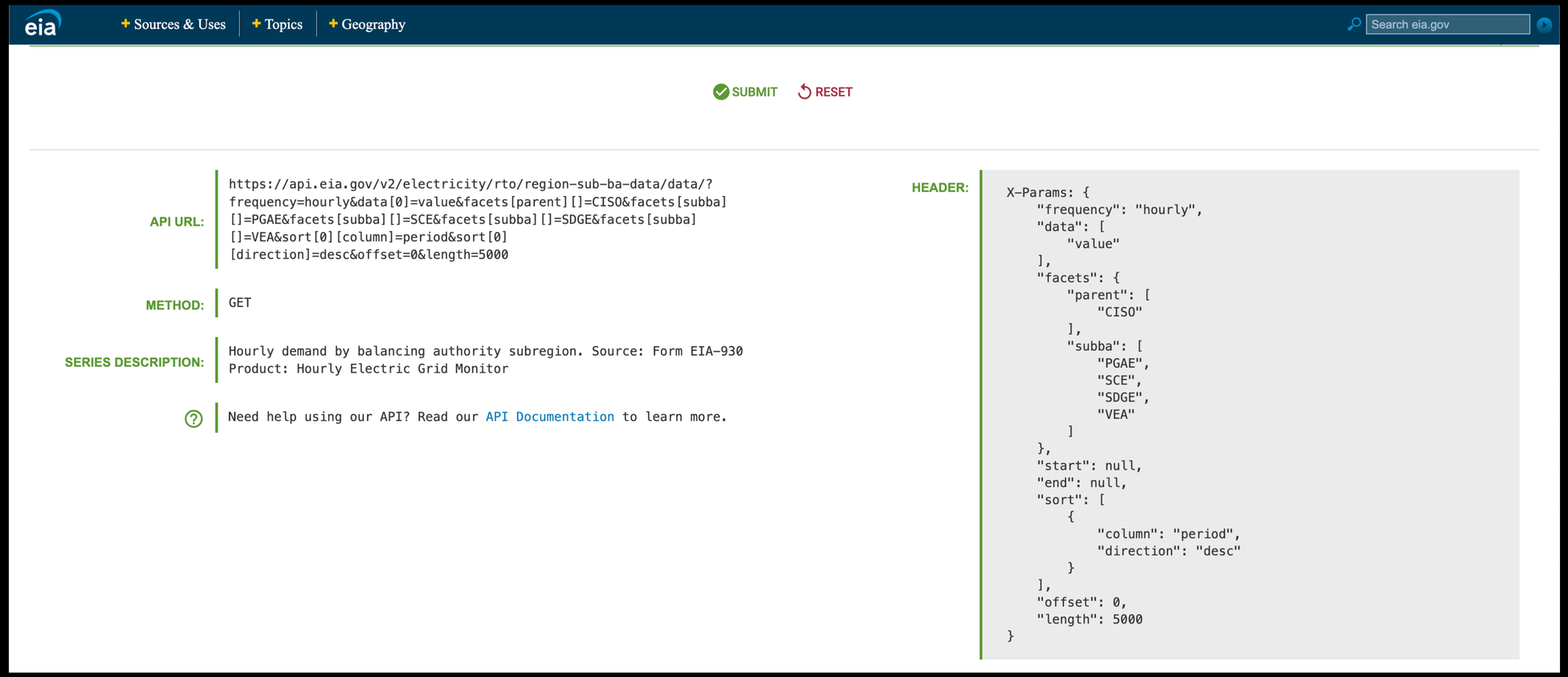
Pipeline Design



Data Source - EIA API



Data Source - EIA API



Pipeline Design

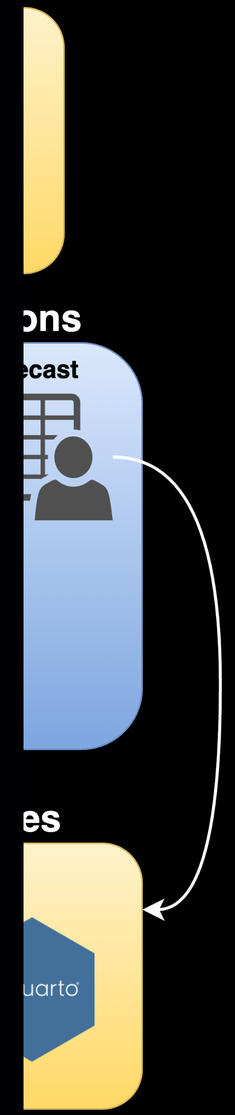
EIA API


```
FROM python:3.10-slim AS builder

ARG QUARTO_VER="1.5.56"
ARG VENV_NAME="my_project"
ENV QUARTO_VER=$QUARTO_VER
ENV VENV_NAME=$VENV_NAME
RUN mkdir requirements

COPY install_requirements.sh requirements/


COPY requirements.txt requirements/
RUN bash ./requirements/install_requirements.sh $VENV_NAME


FROM python:3.10-slim

ARG QUARTO_VER="1.5.56"
ARG VENV_NAME="my_project"
ENV QUARTO_VER=$QUARTO_VER
ENV VENV_NAME=$VENV_NAME

COPY --from=builder /opt/$VENV_NAME /opt/$VENV_NAME

COPY install_requirements.sh install_quarto
RUN bash ./requirements/install_dependencie
RUN bash ./requirements/install_quarto.sh $QUARTO_VER

RUN echo "source /opt/$VENV_NAME/bin/activate" >> ~/.bashrc
```

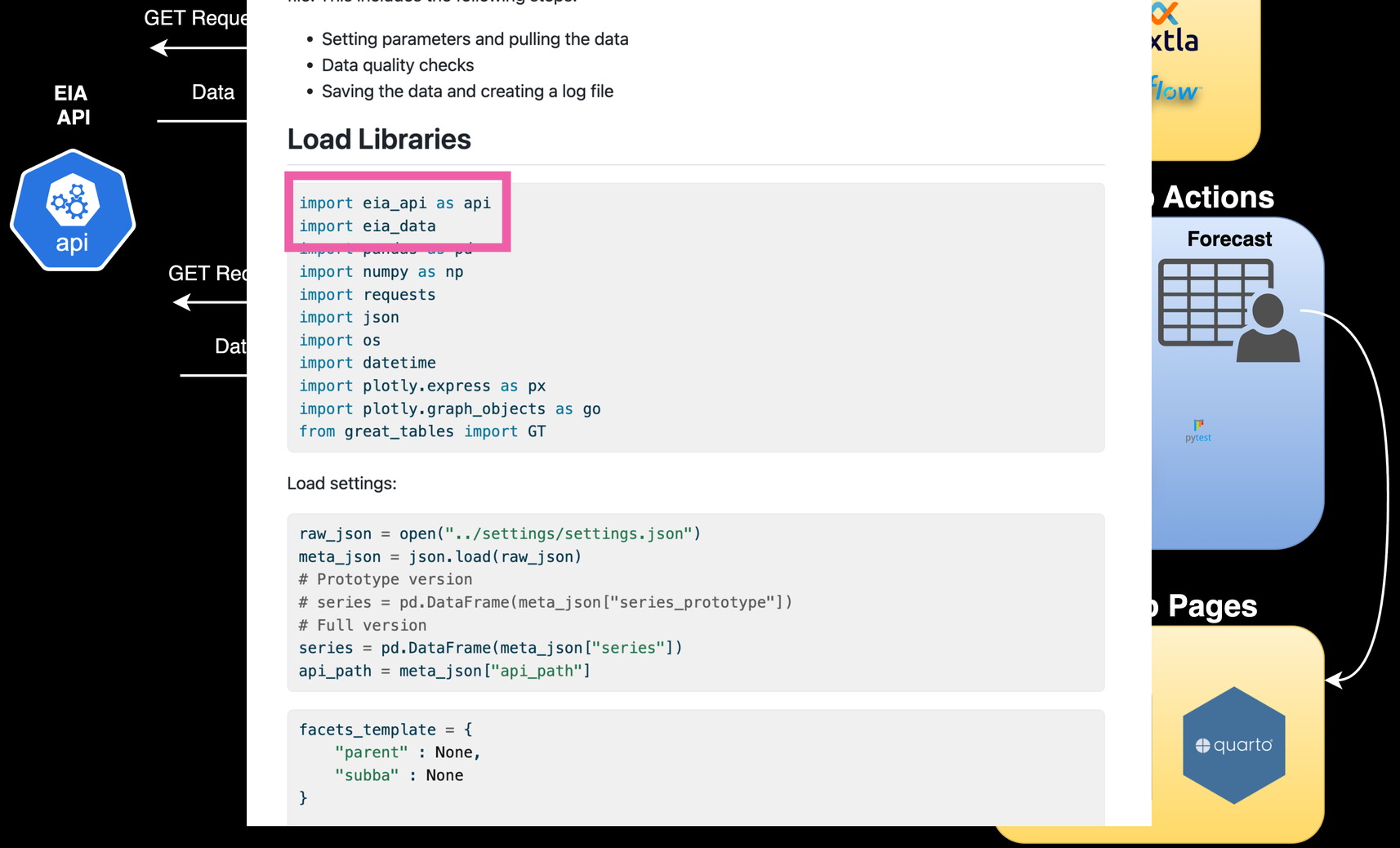
$QUARTO_VER

Pipeline D


```
{
    "name": "PyData NY 2024 Workshop"
    "image": "docker.io/rkrispin/pydata_ny_workshop:arm64.0.0.3",
    // "image": "docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3",
    "customizations" : {
        "vscode": {
            "settings": {
                "python.defaultInterpreterPath": "/opt/pydata-ny-workshop/bin/python3",
                "python.selectInterpreter": "/opt/pydata-ny-workshop/bin/python3"
            },
            "extensions": [
                // Container
                "ms-azuretools.vscode-docker",
                "ms-vscode-remote.remote-containers",
                //Python
                "ms-python.python",
                "ms-python.vscode-pylance",
                // HTML
                "ms-vscode.live-server",
                // Documents
                "yzhang.markdown-all-in-one",
                "quarto.quarto",
                "redhat.vscode-yaml",
                "ms-toolsai.jupyter",
                //Code
                "oderwat.indent-rainbow",
                // Data
                "mechatroner.rainbow-csv",
                "ms-toolsai.datawrangler"
            ]
        }
    },
    "remoteEnv": {
        "EIA_API_KEY": "${localEnv:EIA_API_KEY}"
    }
}
```



Pipeline Design



Data Backfill

The goal of the backfill process is to pull the historical data for the required series using the settings.json file. This includes the following steps:

- Setting parameters and pulling the data
- Data quality checks
- Saving the data and creating a log file

Load Libraries

```
import eia_api as api
import eia_data
import numpy as np
import requests
import json
import os
import datetime
import plotly.express as px
import plotly.graph_objects as go
from great_tables import GT
```

Load settings:

```
raw_json = open("../settings/settings.json")
meta_json = json.load(raw_json)
# Prototype version
# series = pd.DataFrame(meta_json["series_prototype"])
# Full version
series = pd.DataFrame(meta_json["series"])
api_path = meta_json["api_path"]
```

```
facets_template = {
    "parent" : None,
    "subba" : None
}
```

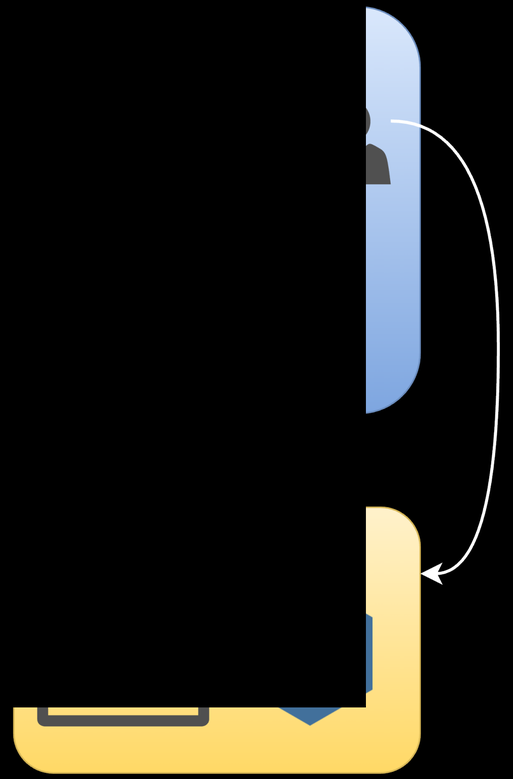
Pipeline Design

Experimentation

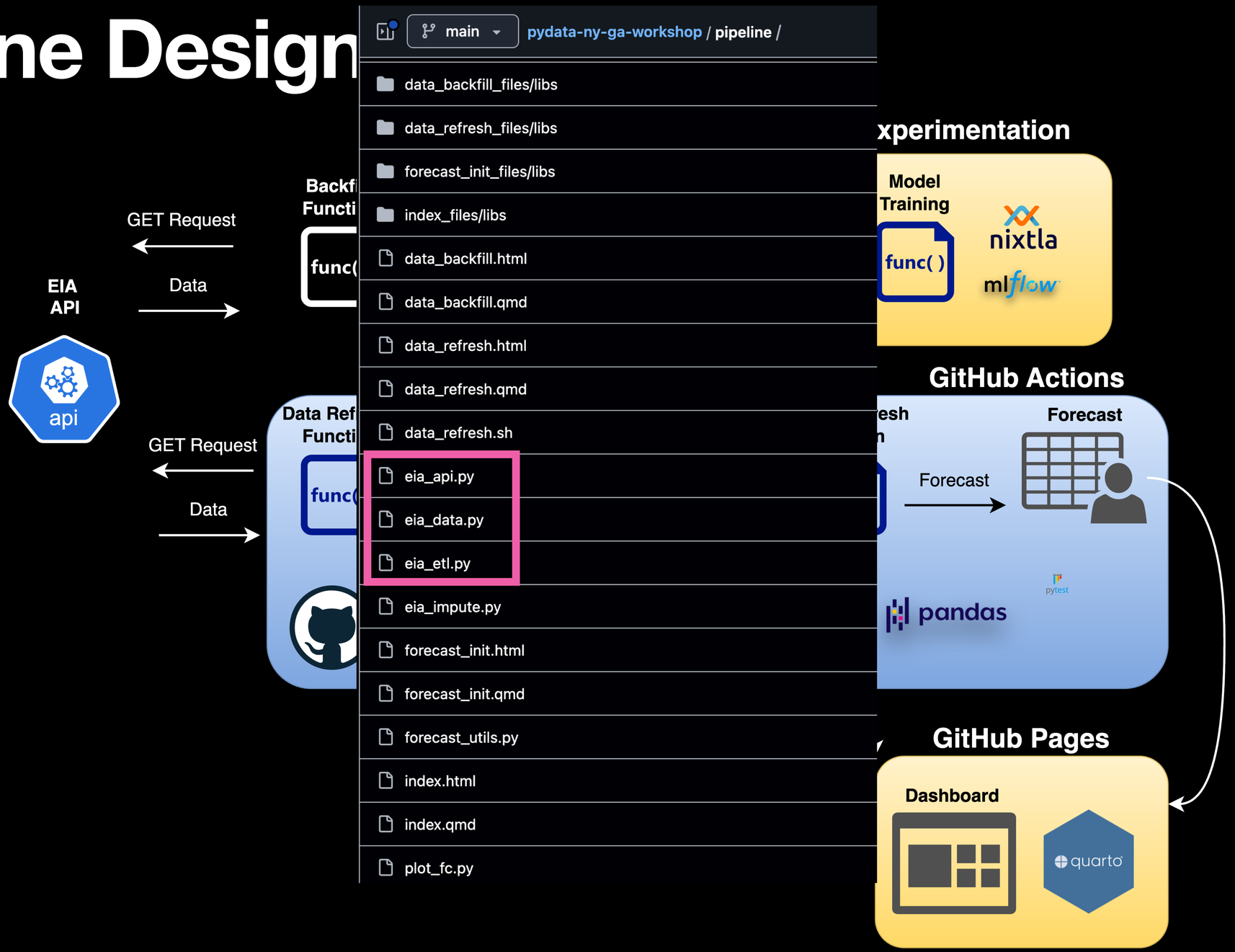
EIA API

/data/data.csv

```
period,subba,impute,y,subba-name,parent,parent-name,value,value-units
2022-01-01 00:00:00,ZONA,,1707.0,,,,,
2022-01-01 01:00:00,ZONA,,1673.0,,,,,
2022-01-01 02:00:00,ZONA,,1644.0,,,,,
2022-01-01 03:00:00,ZONA,,1605.0,,,,,
2022-01-01 04:00:00,ZONA,,1550.0,,,,,
2022-01-01 05:00:00,ZONA,,1487.0,,,,,
2022-01-01 06:00:00,ZONA,,1422.0,,,,,
2022-01-01 07:00:00,ZONA,,1373.0,,,,,
2022-01-01 08:00:00,ZONA,,1336.0,,,,,
2022-01-01 09:00:00,ZONA,,1317.0,,,,,
2022-01-01 10:00:00,ZONA,,1307.0,,,,,
2022-01-01 11:00:00,ZONA,,1315.0,,,,,
2022-01-01 12:00:00,ZONA,,1343.0,,,,,
2022-01-01 13:00:00,ZONA,,1373.0,,,,,
2022-01-01 14:00:00,ZONA,,1395.0,,,,,
2022-01-01 15:00:00,ZONA,,1436.0,,,,,
2022-01-01 16:00:00,ZONA,,1488.0,,,,,
2022-01-01 17:00:00,ZONA,,1526.0,,,,,
2022-01-01 18:00:00,ZONA,,1553.0,,,,,
2022-01-01 19:00:00,ZONA,,1565.0,,,,,
2022-01-01 20:00:00,ZONA,,1574.0,,,,,
2022-01-01 21:00:00,ZONA,,1592.0,,,,,
```



Pipeline Design



Pipeline Design

Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is ava

GitHub Pages

Dashboard

Pipeline Design

Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is ava

GitHub Pages

Dashboard

Pipeline Design

Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is ava

GitHub Pages

Dashboard

Pipeline Design

Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is ava

GitHub Pages

Dashboard

Pipeline Design

Experimentation

parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	imputed	type	update	success	comments
NYIS	ZONA	2024-11-05 05:30:28.064184+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:30:39.101353+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONC	2024-11-05 05:30:54.164038+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZOND	2024-11-05 05:31:09.183913+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONE	2024-11-05 05:31:24.350373+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONF	2024-11-05 05:31:38.829825+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONG	2024-11-05 05:31:54.169900+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONH	2024-11-05 05:32:08.296707+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONI	2024-11-05 05:32:23.408402+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONJ	2024-11-05 05:32:39.046172+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONK	2024-11-05 05:32:56.003404+00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	2022-01-01 00:00:00	2024-10-01 01:00:00	True	True	24098.0	1.0	1	backfill	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:07.265687+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONB	2024-11-05 05:34:08.968823+00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	2024-10-01 02:00:00	2024-11-04 02:00:00	True	True	817.0	2.0	2	refresh	True	True	Missing values wer
NYIS	ZONA	2024-11-05 05:34:55.831717+00:00	2024-11-04 03:00:00	2024-11-04 02:00:00							0	refresh	False	True	No new data is ava

GitHub Pages

Dashboard

Pipeline D

Data Refresh

Load Libraries

```
import eia_api as api
import eia_data
import eia_impute as impute
import pandas as pd
import numpy as np
import requests
import json
import os
import datetime
import plotly.express as px
import plotly.graph_objects as go
from great_tables import GT
```

Settings

Load settings:

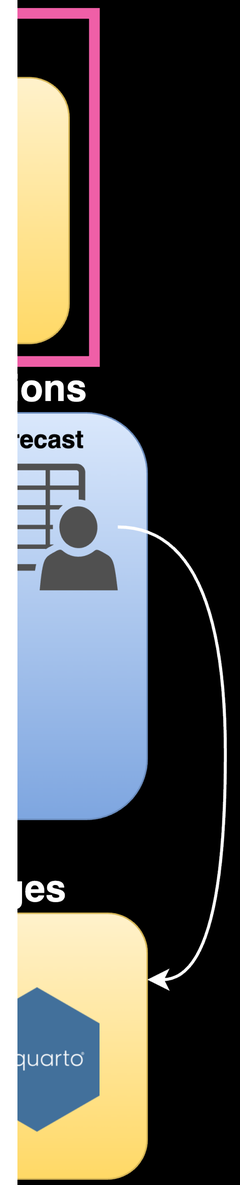
```
raw_json = open("../settings/settings.json")
meta_json = json.load(raw_json)
# Prototype version
# series = pd.DataFrame(meta_json["series_prototype"])
# Full version
series = pd.DataFrame(meta_json["series"])
api_path = meta_json["api_path"]
meta_path = meta_json["meta_path"]
data_path = meta_json["data_path"]
# Should be setting as false during development
save_data = True
save_meta = True

eia_api_key = os.getenv('EIA_API_KEY')
forecast_settings = meta_json["forecast"]
leaderboard_path = meta_json["backtesting"]["leaderboard_path"]
```

Pipeline D

Experimentation

Loading Required Libraries

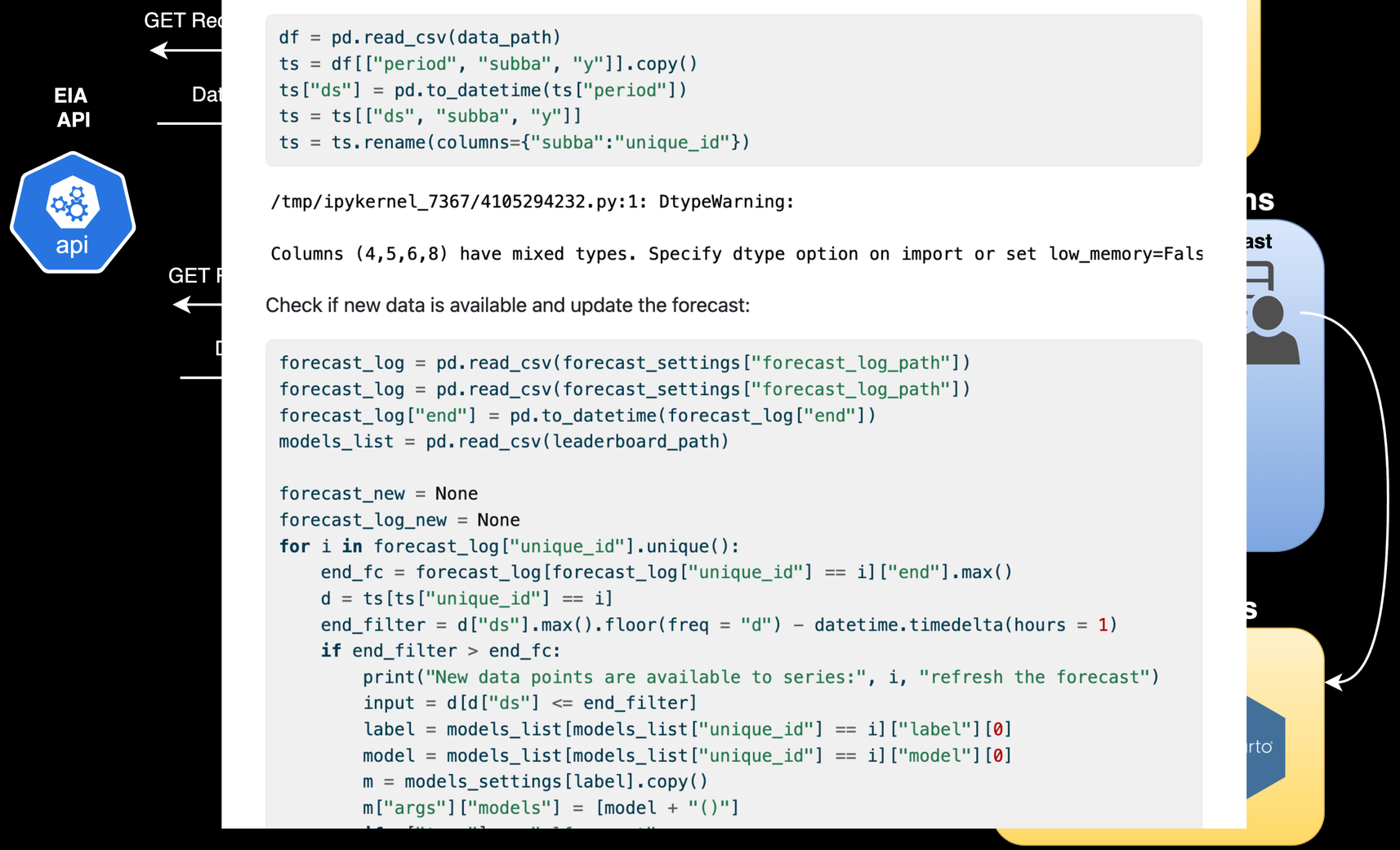
```
import pandas as pd
import numpy as np
import requests
import json
import os
import mlflow
import datetime
import plotly.graph_objects as go
from great_tables import GT


from statsforecast import StatsForecast
from statsforecast.models import (
    HoltWinters,
    CrostonClassic as Croston,
    HistoricAverage,
    DynamicOptimizedTheta,
    SeasonalNaive,
    AutoARIMA,
    AutoETS,
    AutoTBATS,
    MSTL

)

from mlforecast import MLForecast
from mlforecast.target_transforms import Differences
from mlforecast.utils import PredictionIntervals
from window_ops.expanding import expanding_mean
from lightgbm import LGBMRegressor
from xgboost import XGBRegressor
from sklearn.linear_model import LinearRegression
from utilsforecast.plotting import plot_series
from statistics import mean

import backtesting
```



Pipeline Design

model_unique_id	unique_id	label	model	type	partitions	avg_mape	avg_rmse	avg_coverage
model4_LinearRegression	ZONA	model4	LinearRegression	mlforecast	20	0.031451369703253865	63.70007892927969	0.81875
model4_XGBRegressor	ZONB	model4	XGBRegressor	mlforecast	20	0.04990164204689953	61.50748573552928	0.8270833333333334
model4_LinearRegression	ZONC	model4	LinearRegression	mlforecast	20	0.05883017394892269	109.75741267896122	0.8291666666666667
model4_LGBMRegressor	ZOND	model4	LGBMRegressor	mlforecast	20	0.04610411734303354	32.59497249269365	0.8104166666666667
model4_LinearRegression	ZONE	model4	LinearRegression	mlforecast	20	0.0745053032568893	67.76425655956902	0.8125
model4_LinearRegression	ZONF	model4	LinearRegression	mlforecast	20	0.05212821927381432	75.84396938968229	0.8166666666666667
model4_LinearRegression	ZONG	model4	LinearRegression	mlforecast	20	0.048855291350719504	55.6707171450734	0.80625
model4_XGBRegressor	ZONH	model4	XGBRegressor	mlforecast	20	0.05390935709492725	16.710052356332614	0.7854166666666667
model4_XGBRegressor	ZONI	model4	XGBRegressor	mlforecast	20	0.03842874906638785	27.07394224935995	0.7770833333333333
model4_LGBMRegressor	ZONJ	model4	LGBMRegressor	mlforecast	20	0.025910214240455358	163.31549830977897	0.7833333333333333
model4_LinearRegression	ZONK	model4	LinearRegression	mlforecast	20	0.040560537435138336	95.77579962905139	0.7875

Pipeline Design



Forecast Refresh

Reload the data:

```
df = pd.read_csv(data_path)
ts = df[["period", "subba", "y"]].copy()
ts["ds"] = pd.to_datetime(ts["period"])
ts = ts[["ds", "subba", "y"]]
ts = ts.rename(columns={"subba":"unique_id"})
```

```
/tmp/ipykernel_7367/4105294232.py:1: DtypeWarning:

Columns (4,5,6,8) have mixed types. Specify dtype option on import or set low_memory=Fals
```

Check if new data is available and update the forecast:

```
forecast_log = pd.read_csv(forecast_settings["forecast_log_path"])
forecast_log = pd.read_csv(forecast_settings["forecast_log_path"])
forecast_log["end"] = pd.to_datetime(forecast_log["end"])
models_list = pd.read_csv(leaderboard_path)

forecast_new = None
forecast_log_new = None
for i in forecast_log["unique_id"].unique():
    end_fc = forecast_log[forecast_log["unique_id"] == i]["end"].max()
    d = ts[ts["unique_id"] == i]
    end_filter = d["ds"].max().floor(freq = "d") - datetime.timedelta(hours = 1)
    if end_filter > end_fc:
        print("New data points are available to series:", i, "refresh the forecast")
        input = d[d["ds"] <= end_filter]
        label = models_list[models_list["unique_id"] == i]["label"][0]
        model = models_list[models_list["unique_id"] == i]["model"][0]
        m = models_settings[label].copy()
        m["args"]["models"] = [model + "()"]
```

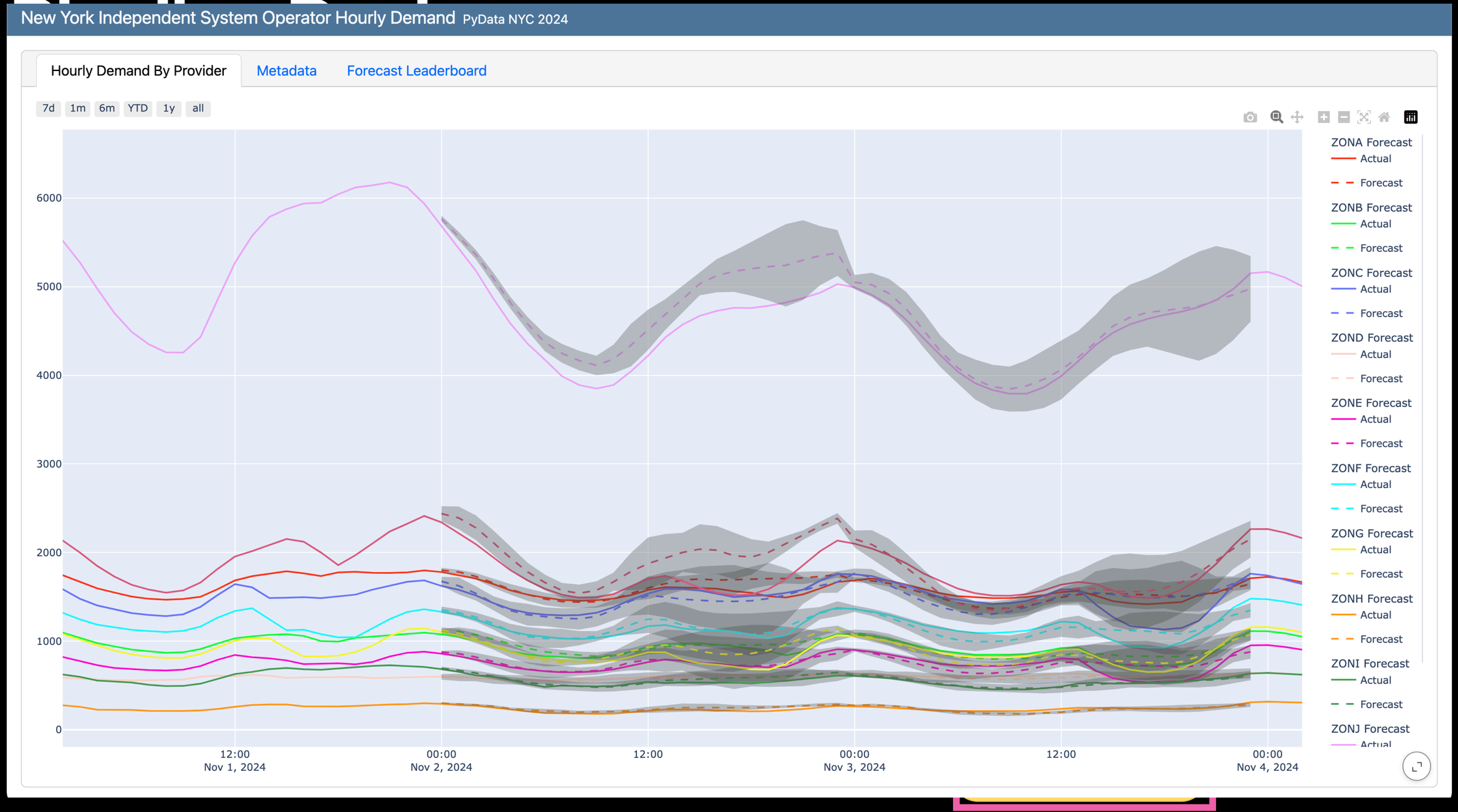
Pipeline Design

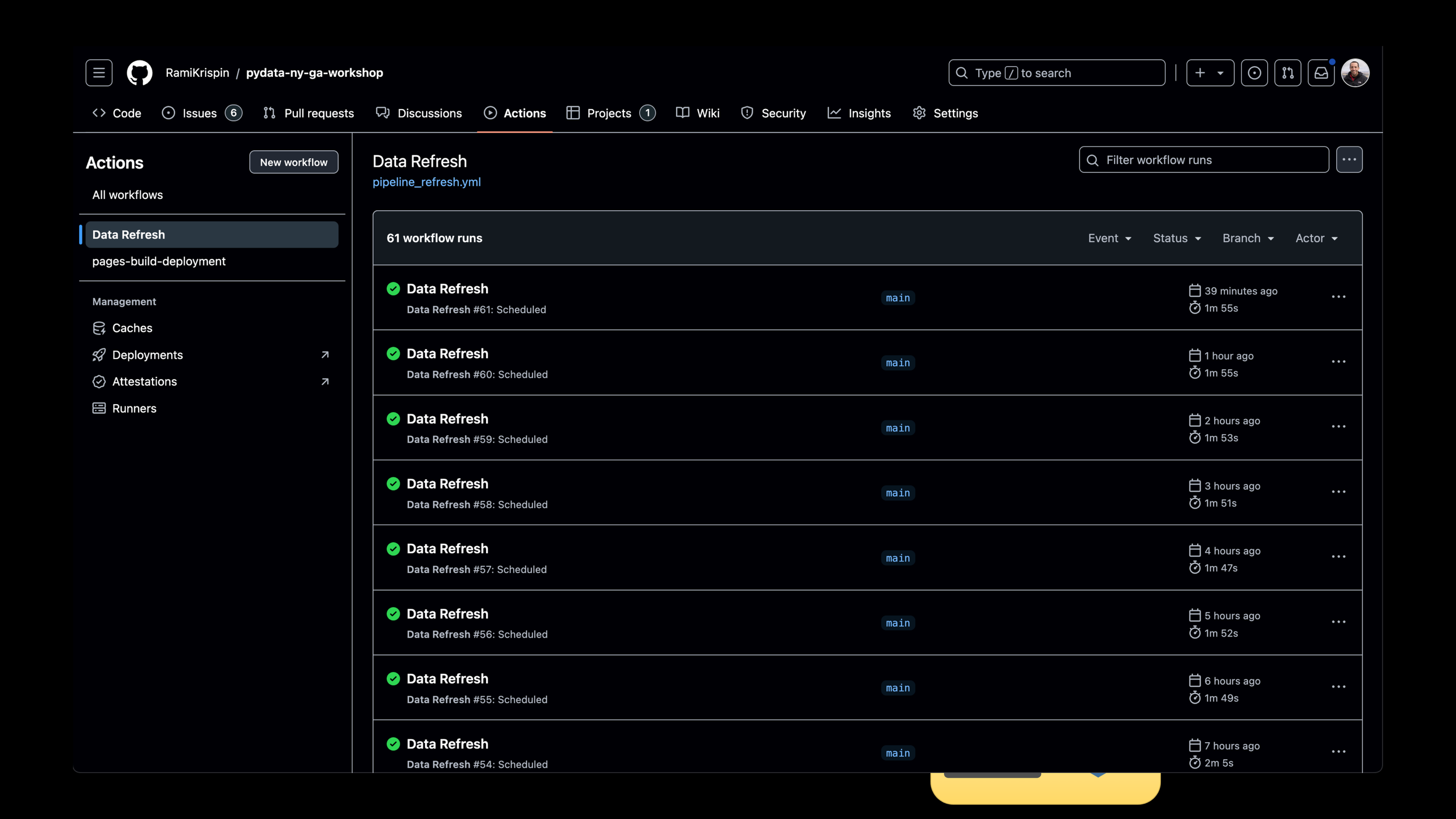
Experimentation

unique_id	ds	forecast	lower	upper	model	label	forecast_label
ZONA	2024-11-02 00:00:00	1800.517800565871	1773.3704623536762	1827.665138778066	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 01:00:00	1780.5180835427157	1755.1184614882377	1805.9177055971936	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 02:00:00	1730.5068915232632	1693.8877826409214	1767.126000405605	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 03:00:00	1654.7358302549685	1603.9235353809265	1705.5481251290105	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 04:00:00	1584.9708637502977	1525.4609120298428	1644.4808154707525	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 05:00:00	1525.9659106867446	1456.1828503187191	1595.74897105477	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 06:00:00	1476.4784068292424	1405.0116680784008	1547.945145580084	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 07:00:00	1447.0144869963324	1381.9366979035979	1512.092276089067	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 08:00:00	1440.470086812012	1376.3634214344702	1504.5767521895536	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 09:00:00	1444.5006873888851	1377.8258236132153	1511.175551164555	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 10:00:00	1477.0962733051292	1410.6160902619663	1543.5764563482921	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 11:00:00	1538.3728728922558	1455.8060075247724	1620.9397382597392	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 12:00:00	1600.5665991741648	1475.2608582003436	1725.872340147986	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 13:00:00	1648.9442846152804	1520.1779561471667	1777.710613083394	LinearRegression	model4	2024-11-01
ZONA	2024-11-02 14:00:00	1688.4020113336267	1572.7048978504054	1804.099124816848	LinearRegression	model4	2024-11-01

Pipeline Design

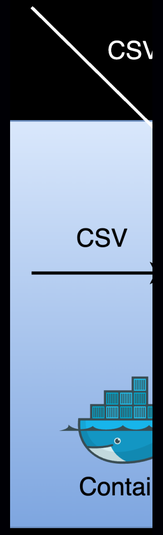
unique_id	model	label	forecast_label	start	end	n_obs	h	refresh_time	success	mape	rmse	coverage	score
ZONA	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:09	True	0.04365224383524807	95.44281545897655	0.7083333333333334	True
ZONB	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:10	True	0.04897028797549269	55.111854724106706	0.6666666666666666	True
ZONC	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:10	True	0.02942917761312673	54.89107179763394	0.9583333333333334	True
ZOND	LGBMRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:11	True	0.022954528805170193	18.555757547214046	1.0	True
ZONE	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:12	True	0.022176471731920985	21.898557451386342	0.9583333333333334	True
ZONF	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:12	True	0.026830806664641407	40.537501009876465	1.0	True
ZONG	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:13	True	0.1035057362451028	105.17885990069209	0.75	True
ZONH	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:14	True	0.1064112030516249	25.454014229573012	0.4583333333333333	True
ZONI	XGBRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:15	True	0.05056286468047634	33.89463797766538	0.8333333333333334	True
ZONJ	LGBMRegressor	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:16	True	0.06566055892878288	314.97418947984215	0.08333333333333333	True
ZONK	LinearRegression	model4	2024-11-01	2024-11-02	2024-11-02 23:00:00	24	24	06/11/2024 07:55:16	True	0.13673648504648134	270.2176330551692	0.3333333333333333	True
ZONA	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:17	True	0.04069130036037995	71.60527947080918	0.7083333333333334	True
ZONB	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:18	True	0.06327243341515874	60.72460597309195	0.75	True
ZONC	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:18	True	0.10937874297241784	180.95626539971647	0.7083333333333334	True
ZOND	LGBMRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:19	True	0.030088697473674635	23.02897283243766	0.9166666666666666	True
ZONE	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:19	True	0.12608821062538283	97.36083704095965	0.75	True
ZONF	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:20	True	0.07354674230651301	93.86685356486336	0.7916666666666666	True
ZONG	LinearRegression	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:20	True	0.07126757093240858	65.10305390846375	0.7083333333333334	True
ZONH	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:21	True	0.07273225940792014	20.817034743164285	0.5	True
ZONI	XGBRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:23	True	0.02257437803721225	13.925164828678495	1.0	True
ZONJ	LGBMRegressor	model4	2024-11-02	2024-11-03	2024-11-03 23:00:00	24	24	06/11/2024 07:55:23	True	0.015217619426924294	78.30077282677799	1.0	True


New York Independent System Operator Hourly Demand PyData NYC 2024
Hourly Demand By Provider
Metadata
Forecast Leaderboard
7d
1m
6m
YTD
1y
all
6000
5000
4000
3000
2000
1000
0
12:00
Nov 1, 2024
00:00
Nov 2, 2024
12:00
00:00
Nov 3, 2024
12:00
00:00
Nov 4, 2024
ZONA Forecast
Actual
Forecast
ZONB Forecast
Actual
Forecast
ZONC Forecast
Actual
Forecast
ZOND Forecast
Actual
Forecast
ZONE Forecast
Actual
Forecast
ZONF Forecast
Actual
Forecast
ZONG Forecast
Actual
Forecast
ZONH Forecast
Actual
Forecast
ZONI Forecast
Actual
Forecast
ZONJ Forecast
Actual


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
Data Refresh
pipeline_refresh.yml
Filter workflow runs
61 workflow runs
Event
Status
Branch
Actor
Data Refresh
Data Refresh #61: Scheduled
main
39 minutes ago
1m 55s
Data Refresh
Data Refresh #60: Scheduled
main
1 hour ago
1m 55s
Data Refresh
Data Refresh #59: Scheduled
main
2 hours ago
1m 53s
Data Refresh
Data Refresh #58: Scheduled
main
3 hours ago
1m 51s
Data Refresh
Data Refresh #57: Scheduled
main
4 hours ago
1m 47s
Data Refresh
Data Refresh #56: Scheduled
main
5 hours ago
1m 52s
Data Refresh
Data Refresh #55: Scheduled
main
6 hours ago
1m 49s
Data Refresh
Data Refresh #54: Scheduled
main
7 hours ago
2m 5s

Pipeline Design


```
"series": [
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZONA",
        "subba_name": "West - NYIS"
    },
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZONB",
        "subba_name": "Genesee - NYIS"
    },
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZONC",
        "subba_name": "Central - NYIS"
    },
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZOND",
        "subba_name": "North - NYIS"
    },
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZONE",
        "subba_name": "Mohawk Valley - NYIS"
    },
    {
        "parent_id": "NYIS",
        "parent_name": "New York Independent System Operator",
        "subba_id": "ZONF",
        "subba_name": "Capital - NYIS"
    },
```


```
"model3": {
    "label": "model3",
    "type": "mlforecast",
    "args": {
        "models": [
            "LGBMRegressor()",
            "XGBRegressor()",
            "LinearRegression()"
        ],
        "lags": [
            1,
            24,
            48
        ],
        "date_features": [
            "day",
            "dayofweek",
            "hour"
        ],
        "n_windows": 5,
        "freq": "h",
        "comments": "ML models with seasonal features and short 3 lags (1, 24, 48)"
    }
},
```


Deployment

GitHub Actions Requirements

- Workflow file
- Environment
- Script to execute
- Trigger

```yaml
name: Data Refresh
```

Data Refresh

pipeline_refresh.yml

Filter workflow runs

63 workflow runs | Event | Status | Branch | Actor

Data Refresh Data Refresh #63: Scheduled	main	1 minute ago *In progress*	...
Data Refresh Data Refresh #62: Scheduled	main	1 hour ago 2m 4s	...
Data Refresh Data Refresh #61: Scheduled	main	1 hour ago 1m 55s	...
Data Refresh Data Refresh #60: Scheduled	main	2 hours ago 1m 55s	...
Data Refresh Data Refresh #59: Scheduled	main	4 hours ago 1m 53s	...
Data Refresh Data Refresh #58: Scheduled	main	5 hours ago 1m 51s	...
Data Refresh Data Refresh #57: Scheduled	main	6 hours ago 1m 47s	...
Data Refresh Data Refresh #56: Scheduled	main	7 hours ago 1m 52s	...

```yaml
USER_NAME: ${{ secrets.USER_NAME }}
```

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

```yaml
name: Data Refresh

on:
  schedule:
    - cron: "0 */1 * * *"
jobs:
  refresh-the-dashboard:
    runs-on: ubuntu-22.04
    container:
      image: docker.io/rkrispin/pydata_ny_workshop:amd64.0.0.3
    steps:
      - name: checkout_repo
        uses: actions/checkout@v3
        with:
          ref: "main"
      - name: Data Refresh
        run: bash ./pipeline/data_refresh.sh
        env:
          EIA_API_KEY: ${{ secrets.EIA_API_KEY }}
          USER_EMAIL: ${{ secrets.USER_EMAIL }}
          USER_NAME: ${{ secrets.USER_NAME }}
```

Unit Tests!!!

What else can we do?

Many other cool things!

Use Cases

- Alerts
- Social
- Data visualization
- Automation

Best Practices

- Set prototype version
- Know the tools limitations
- Using settings files and environment variables
- Unit tests
- Interactive documents
- GitHub Templates!

Summary

- GitHub Actions enables to schedule workflows
- Docker provides a reproducible environment
- Great use cases - open source projects
- Quarto documents to host the pipeline
- Leverage great open source libraries (MLflow, Y-Data Profile, etc.)
- Not enterprise ready

Questions?

Get In Touch