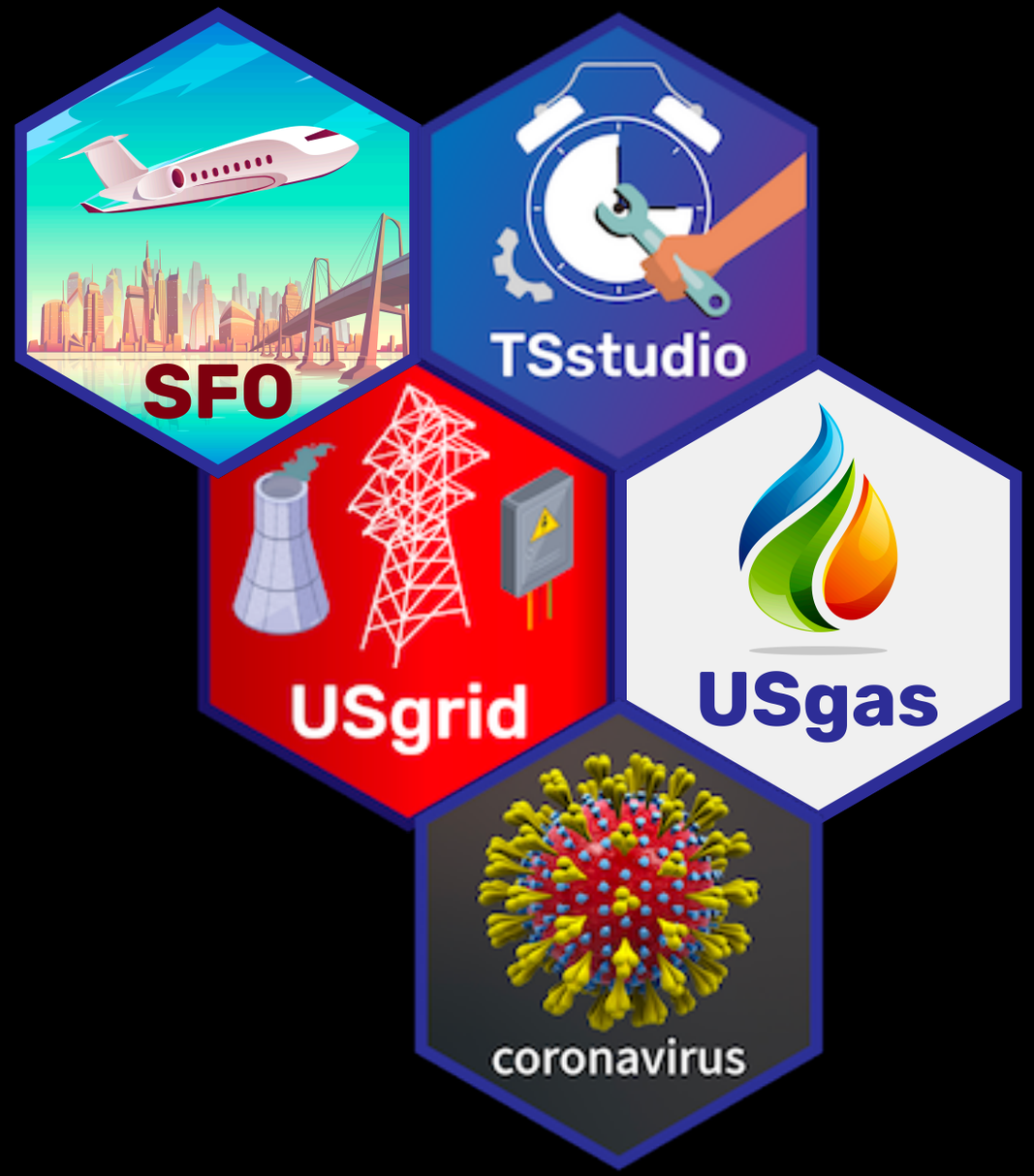
Deploy and Monitor ML Pipelines with Open Source and Free Tools

AI_Dev Conference

Rami Krispin, June 19th, 2024

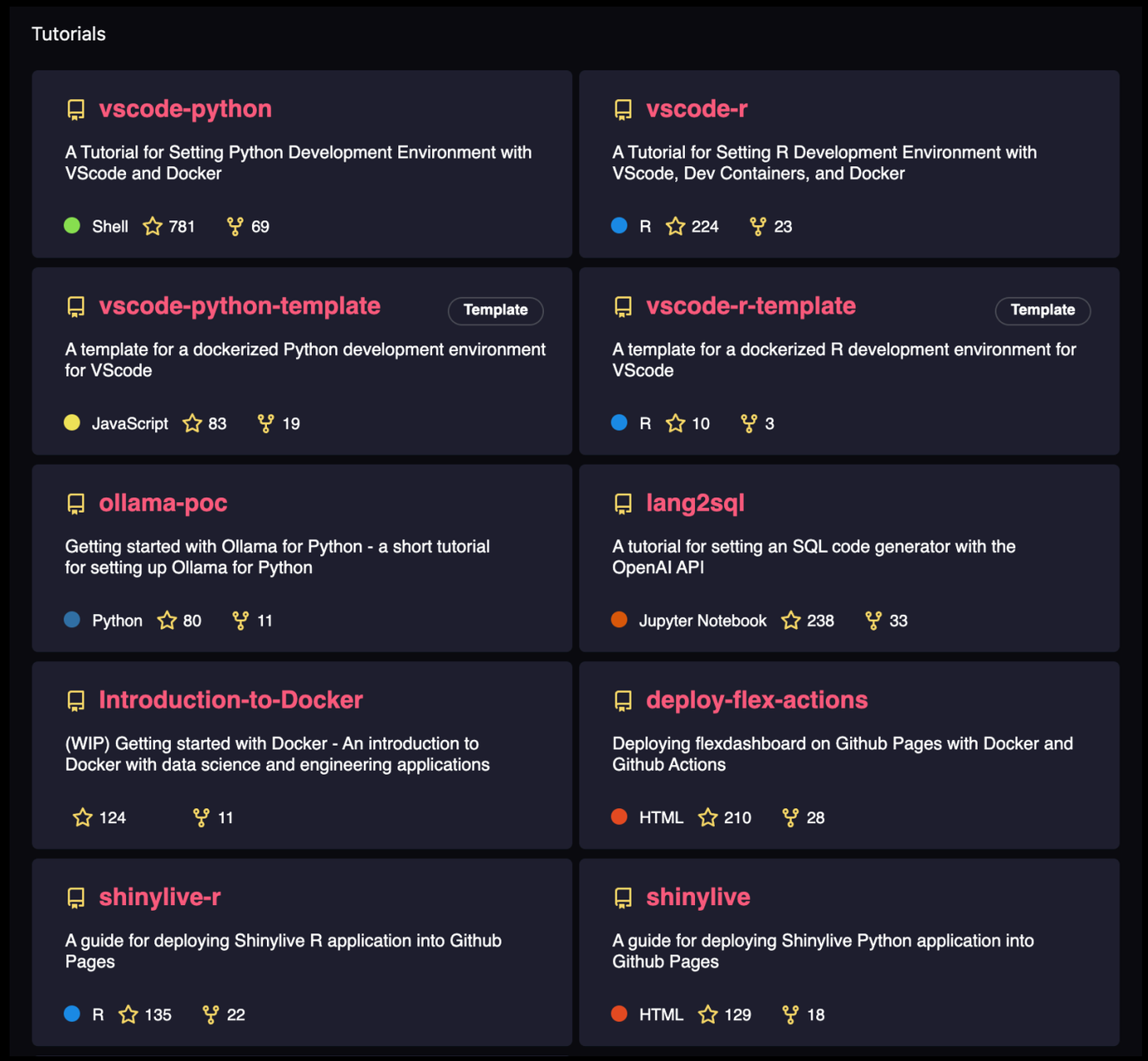
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author



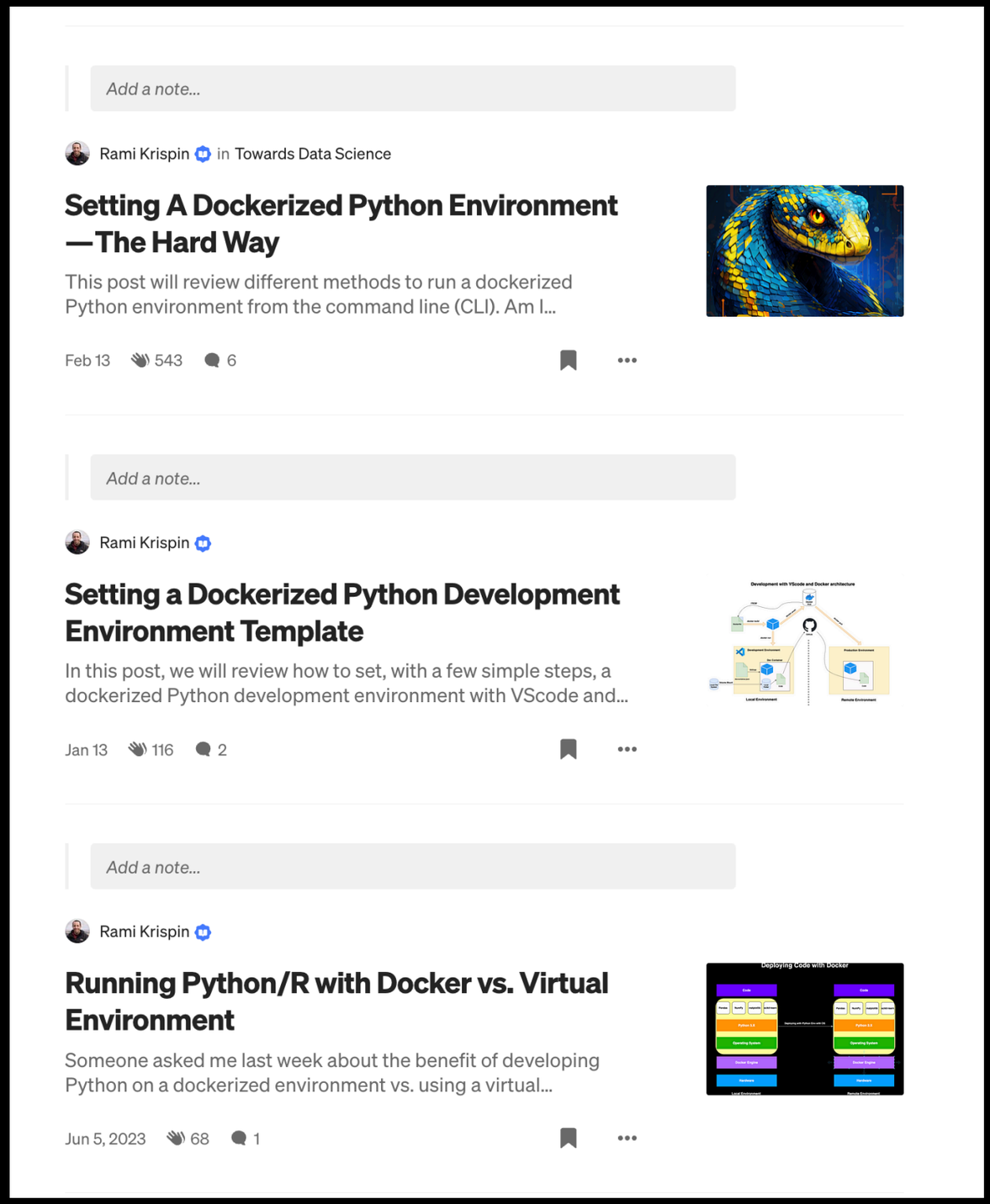
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author



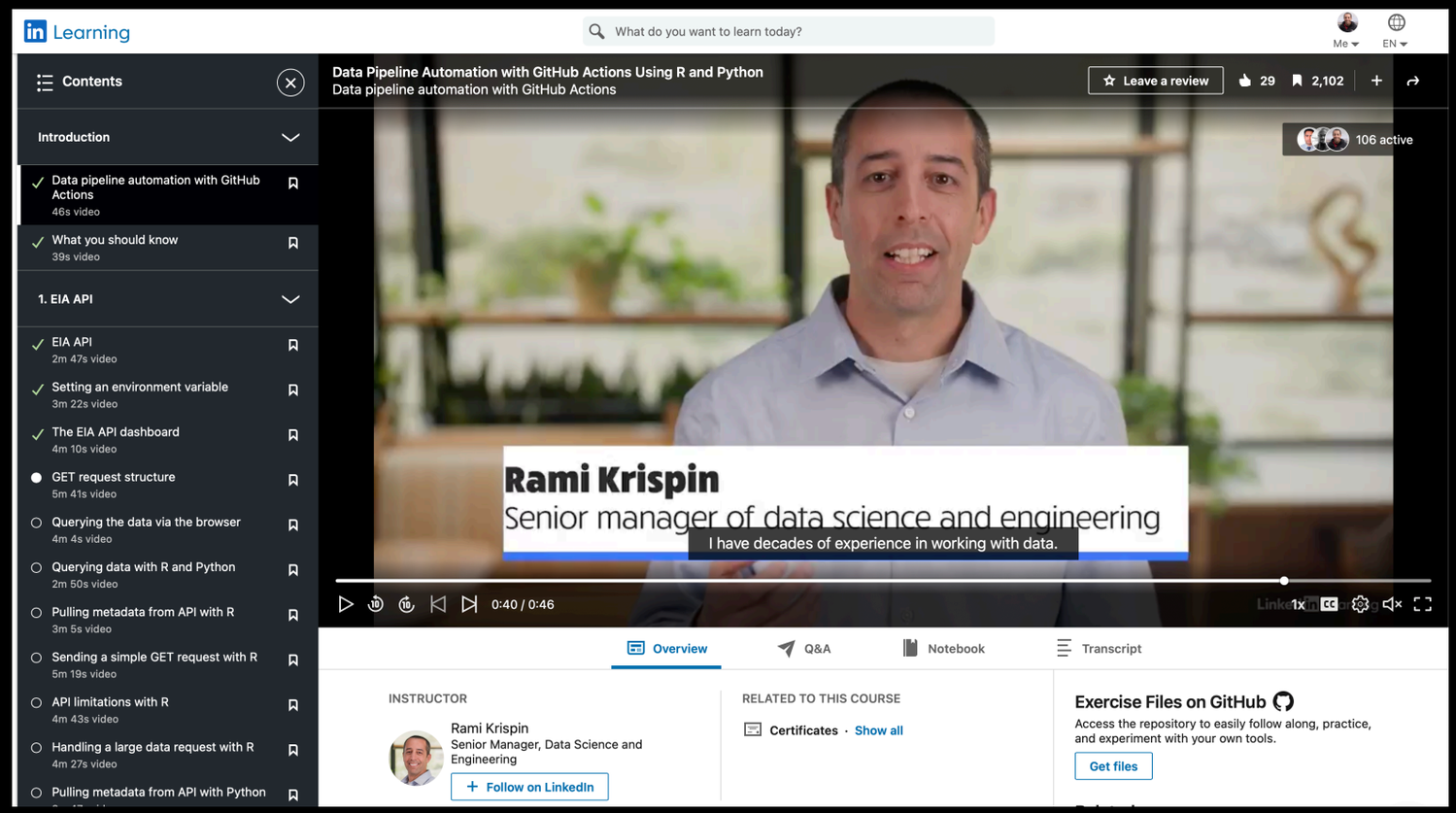
About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author



About Me

- Senior Manager
- Forecasting
- MLOps
- Open Source
- Author



About Me

- Senior Manager
- **Forecasting**
- **MLOps**
- **Open Source**
- Author

https://github.com/RamiKrispin/ai-dev-2024-ml-workshop

Poll

Are You Familiar with?

- Docker
- GitHub Actions
- Quarto
- MLflow
- Python/R

Docker

What is Docker?

Accelerate how you build, share, and run applications

GitHub Actions

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

Features Actions Packages Security Codespaces Copilot Code review Search Issues Discussions

GitHub Actions

Automate your workflow from idea to production

GitHub Actions makes it easy to automate all your software workflows, now with world-class CI/CD. Build, test, and deploy your code right from GitHub. Make code reviews, branch management, and issue triaging work the way you want.

Get started with Actions >

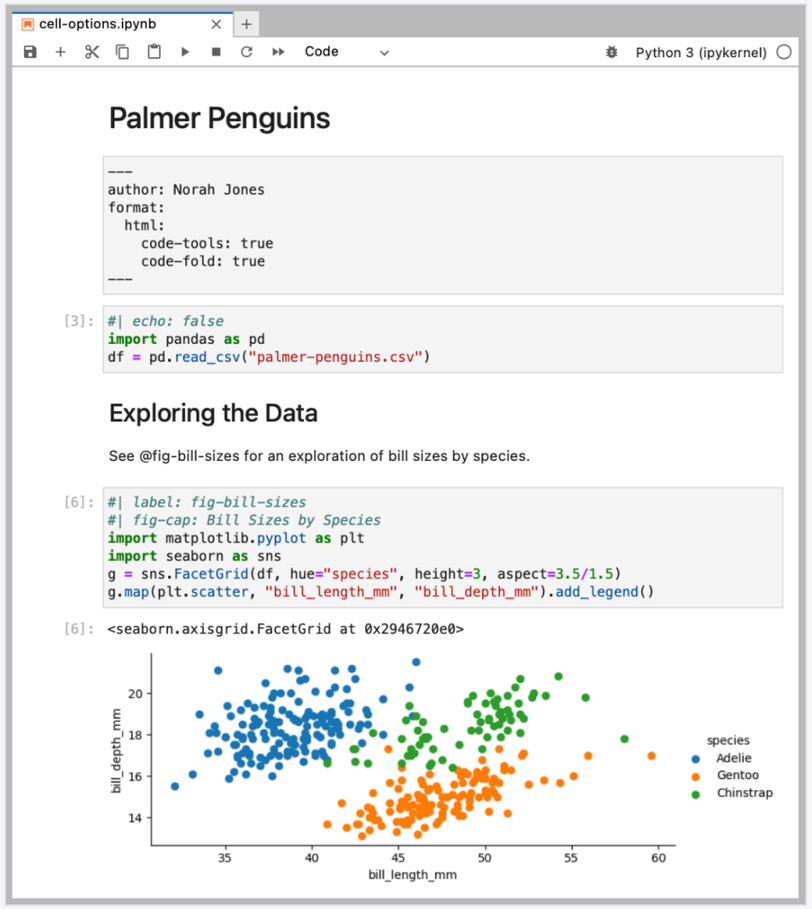
Questions? **Contact Sales >**

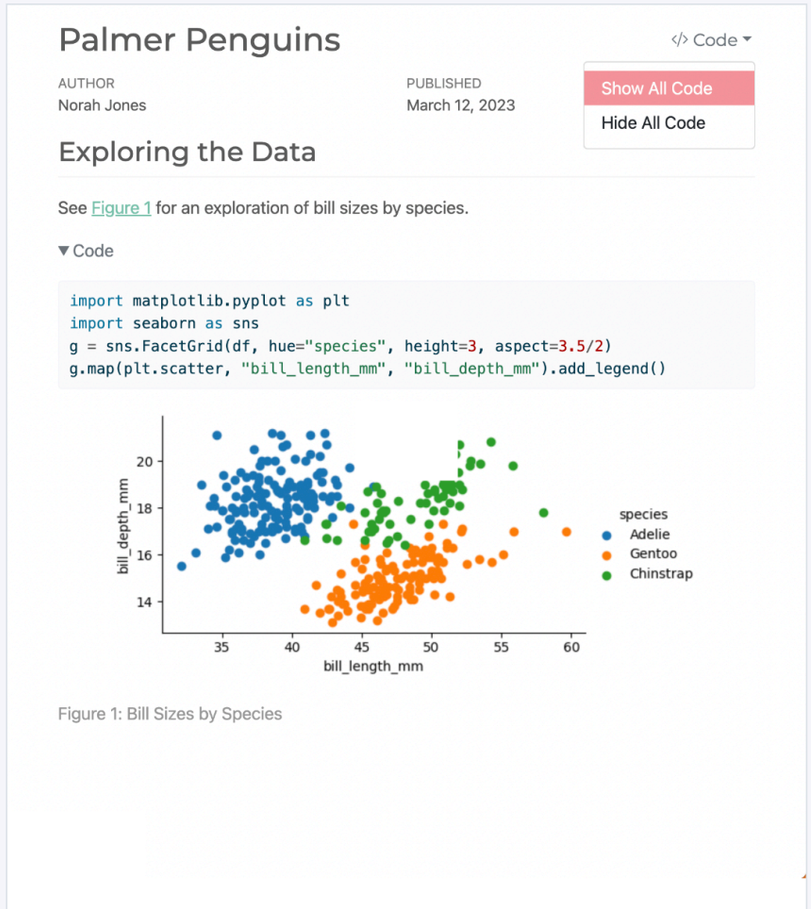
Quarto

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

MLflow

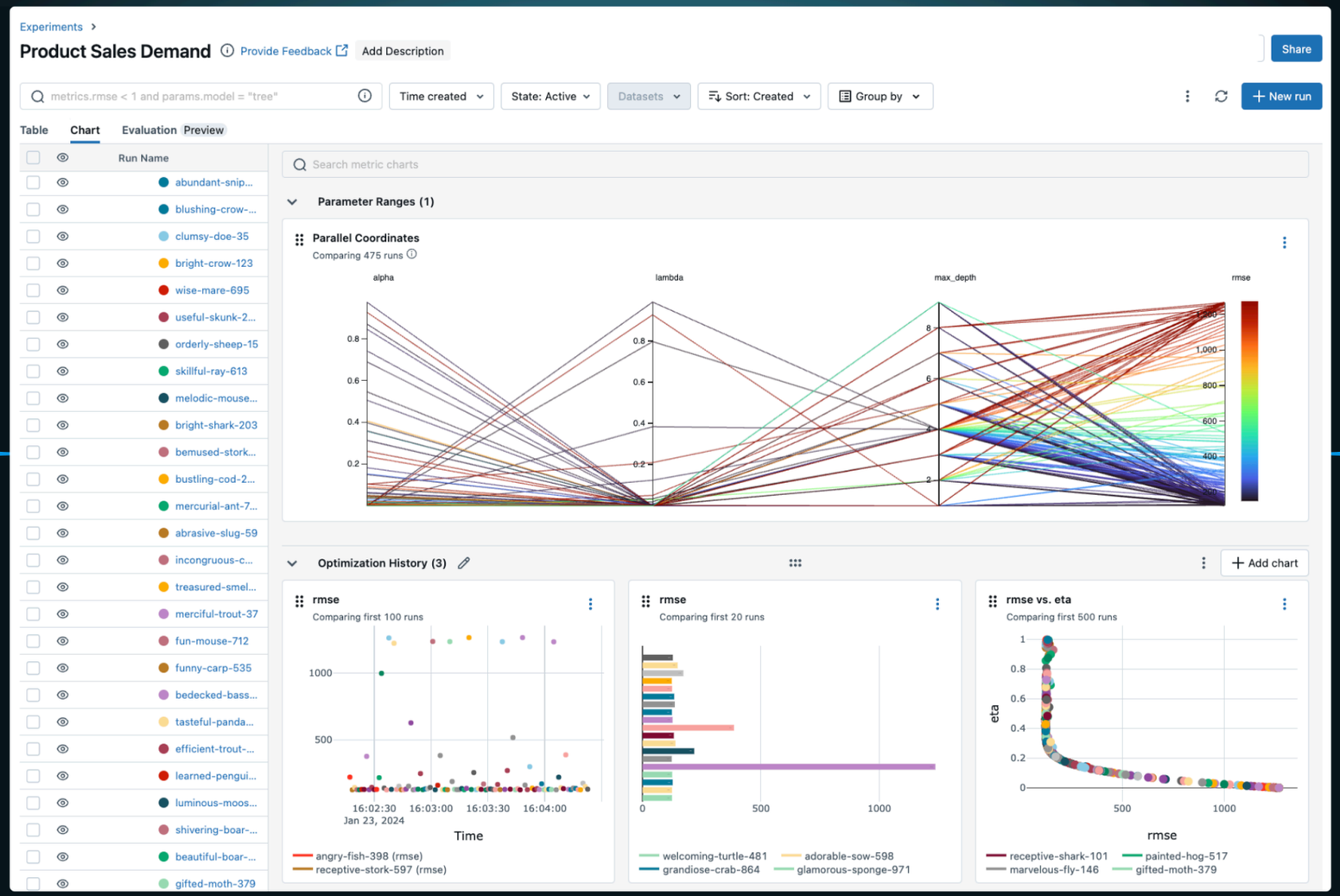
mlflow

Blog Releases Docs Contribute Ambassador Program

Get Started

Get Started

v2.14.0 See the latest release



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

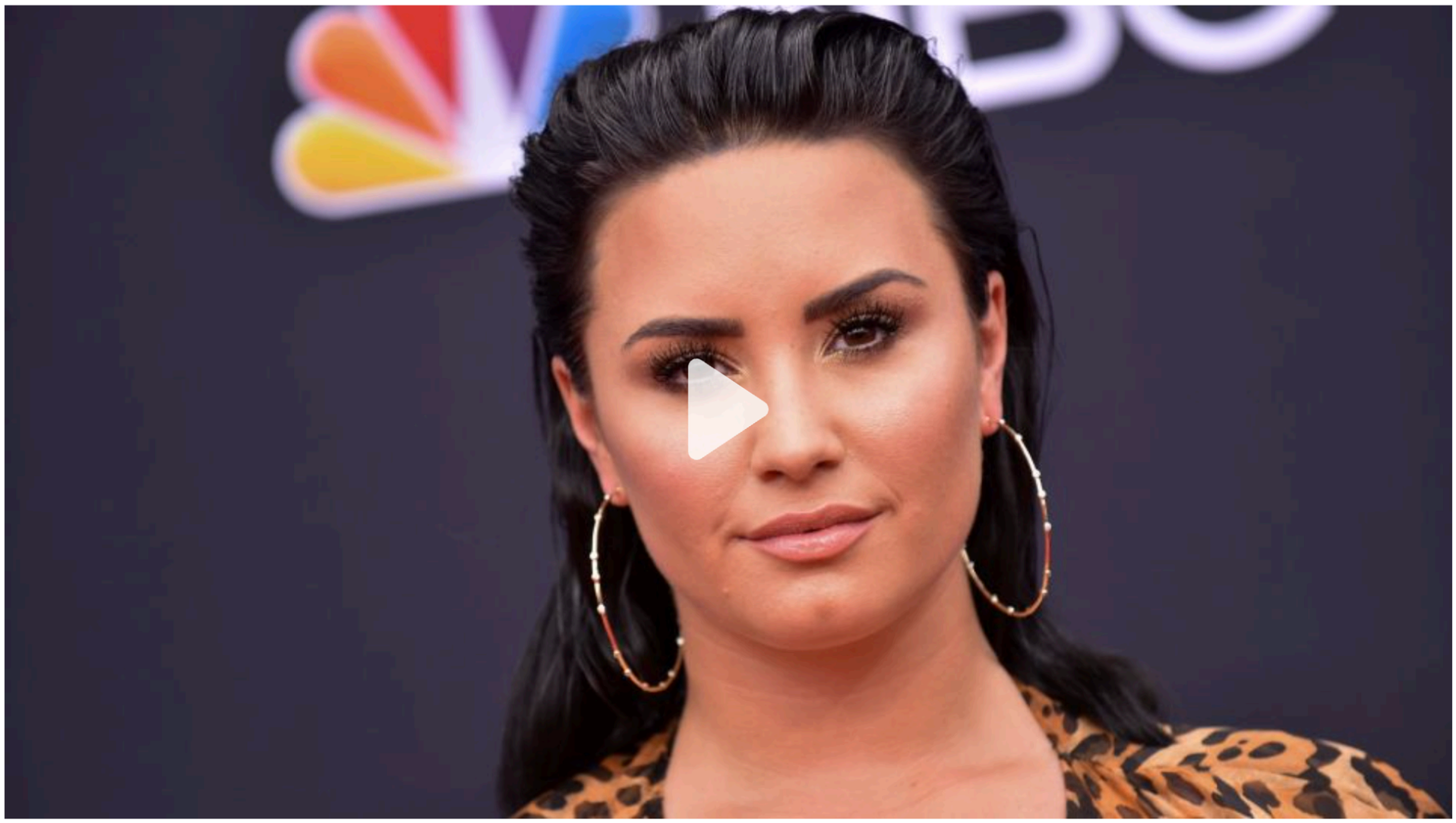
Jennifer Lopez's epic Super Bowl flag coat was custom Versace

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

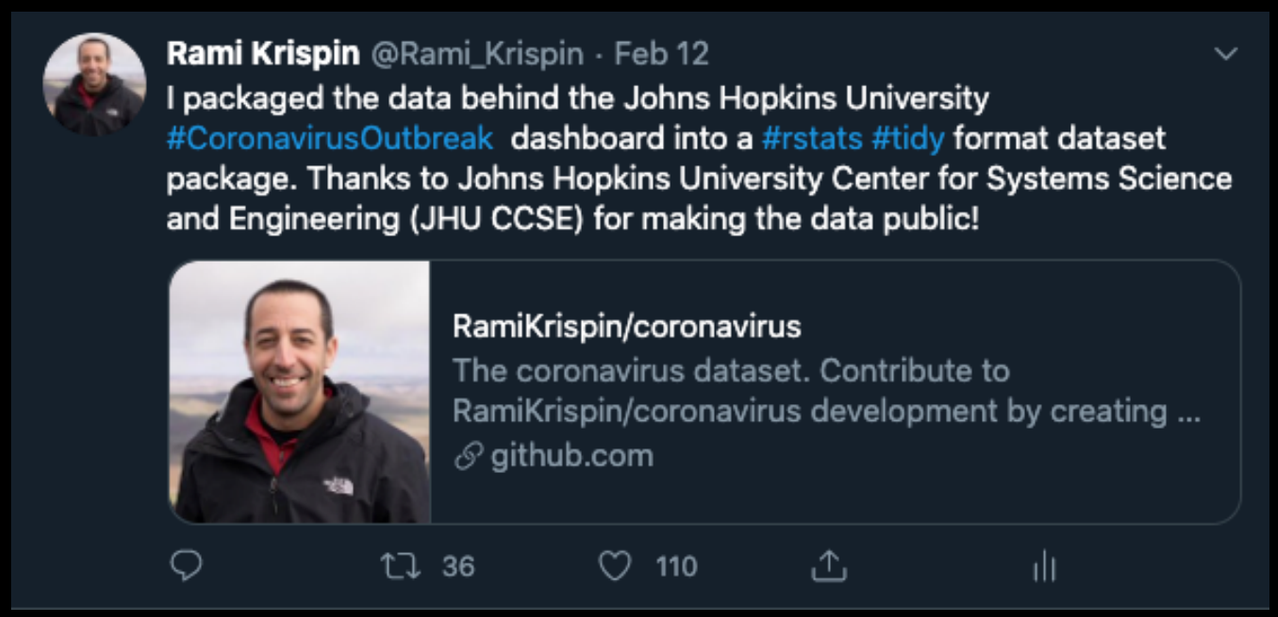
Updated 6:21 PM EST, Sun February 2, 2020

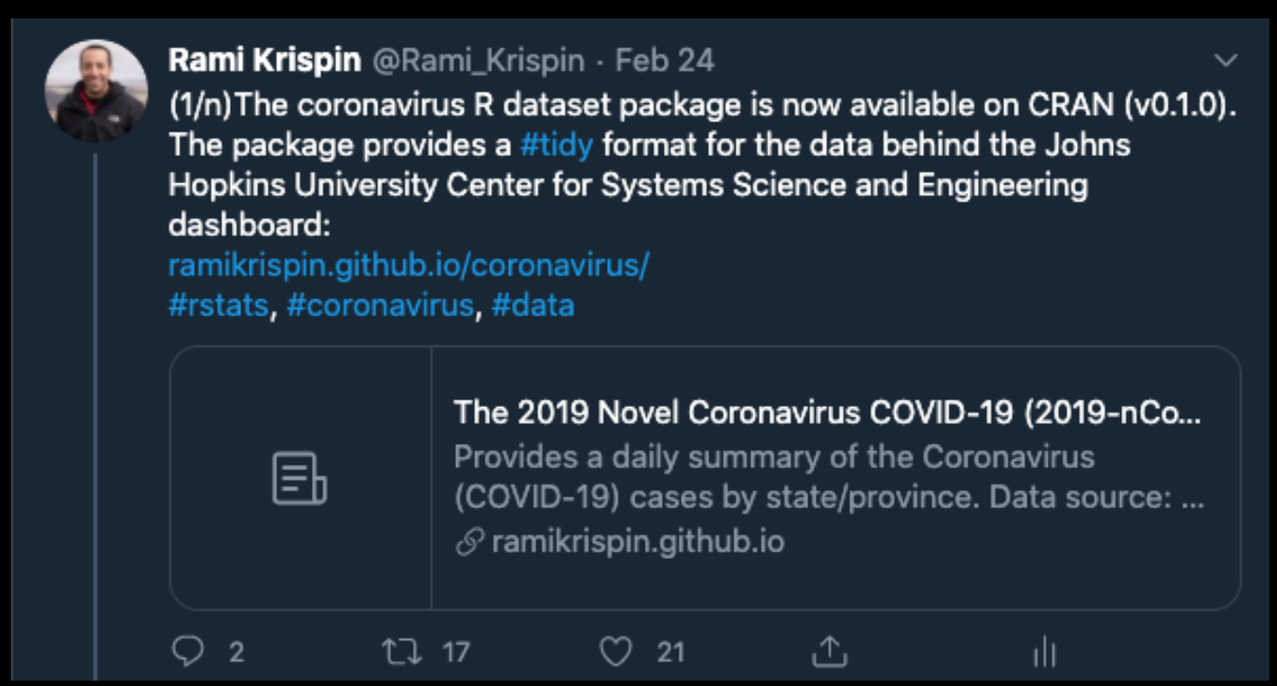
Source:
CNN

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

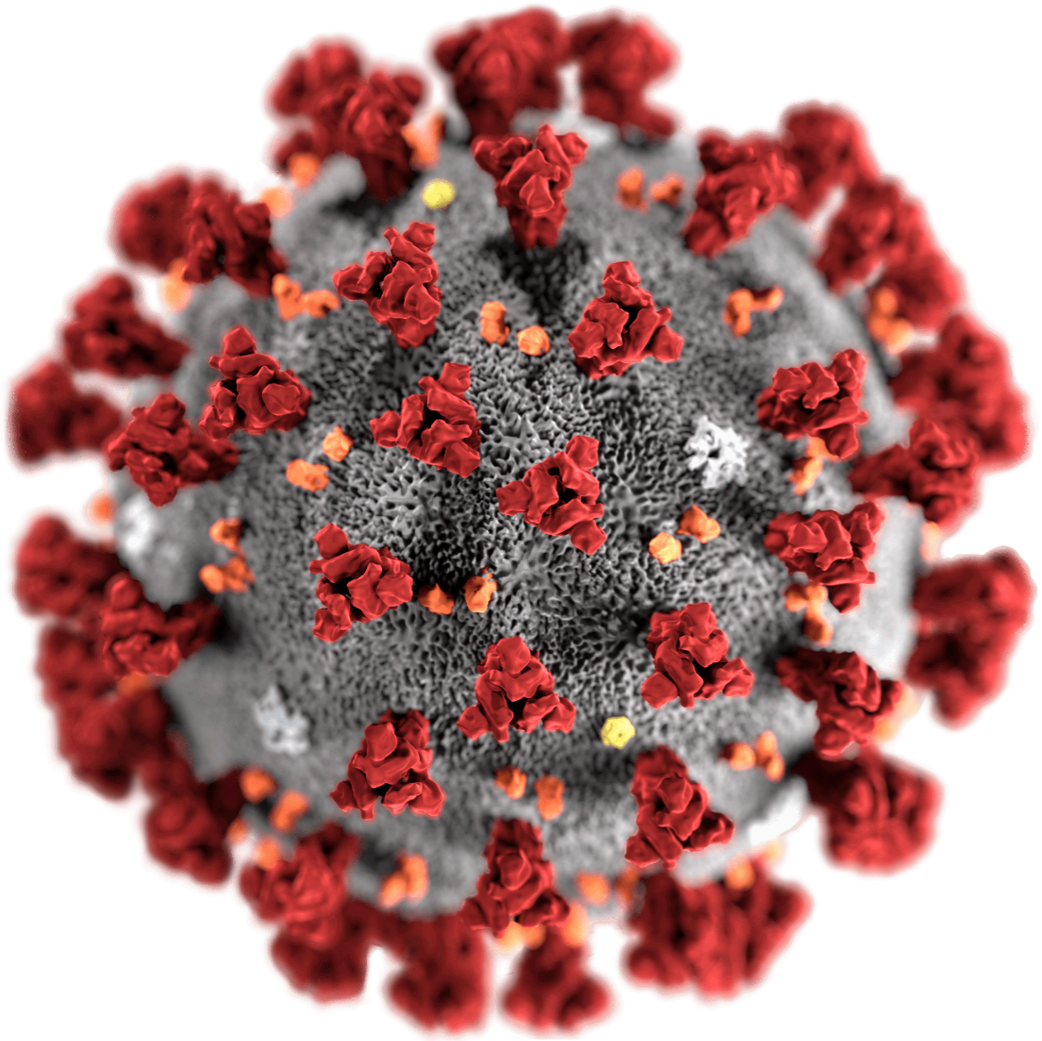

Rami Krispin @Rami_Krispin · Feb 24
(1/n)The coronavirus R dataset package is now available on CRAN (v0.1.0). The package provides a #tidy format for the data behind the Johns Hopkins University Center for Systems Science and Engineering dashboard:
ramikrispin.github.io/coronavirus/
#rstats, #coronavirus, #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

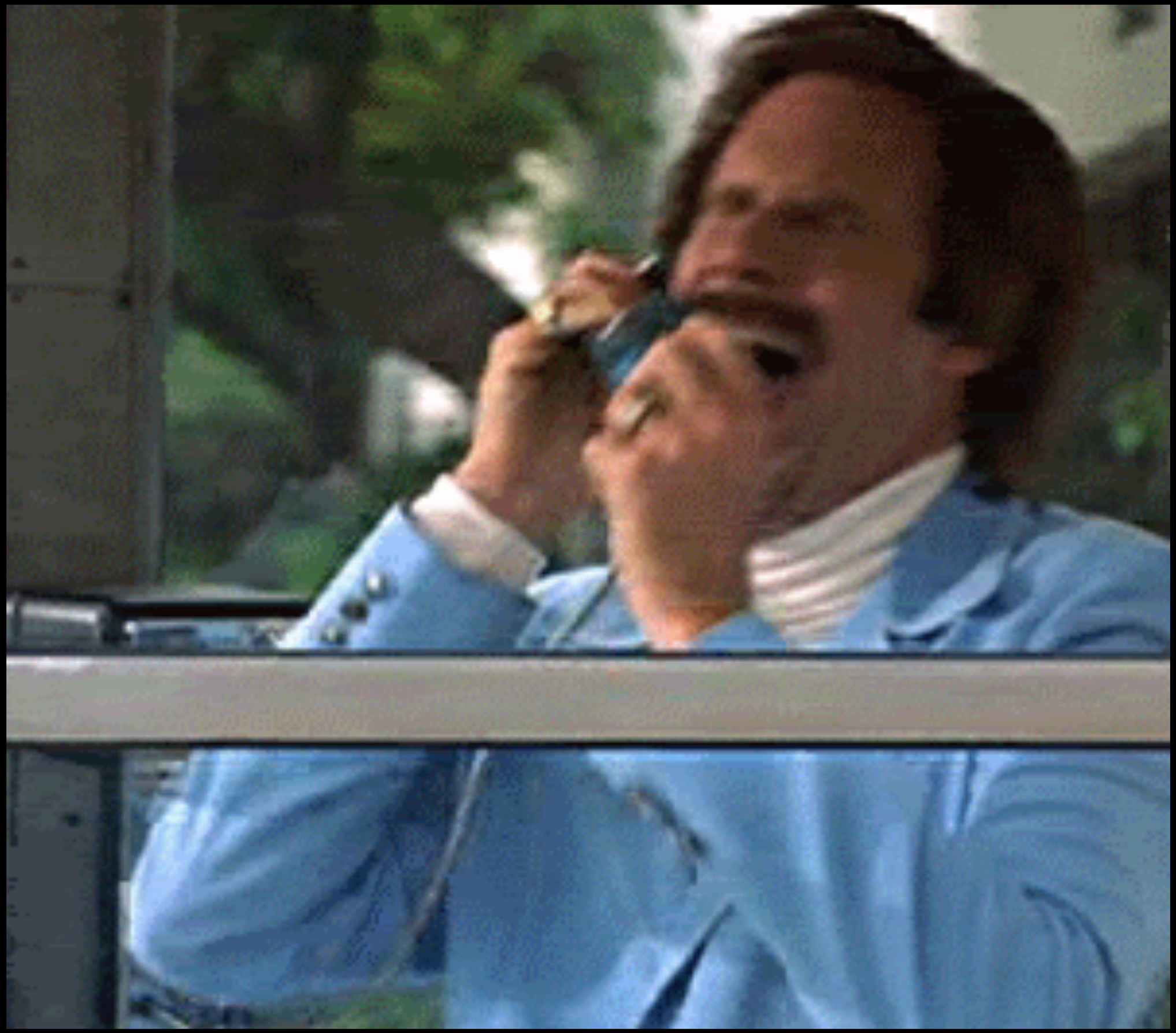
Negative case values — 2
#5 by j4yr0u93 was closed on Mar 10, 2020

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

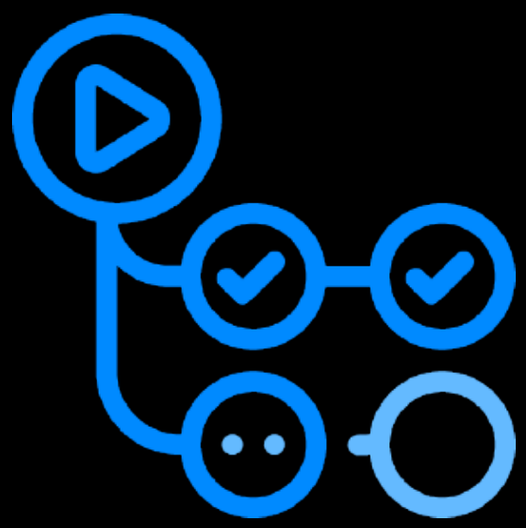
Adding a country filter to the dashboard — 3
#3 by Agusum was closed on Mar 8, 2020

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

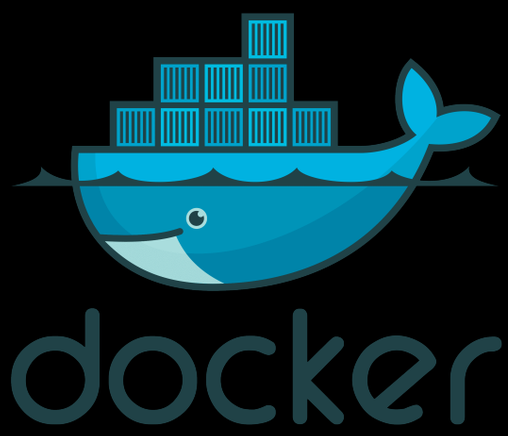
is there any plan to automate data update? — 16
#1 by statkclee was closed on Apr 25, 2020



Actions + Docker

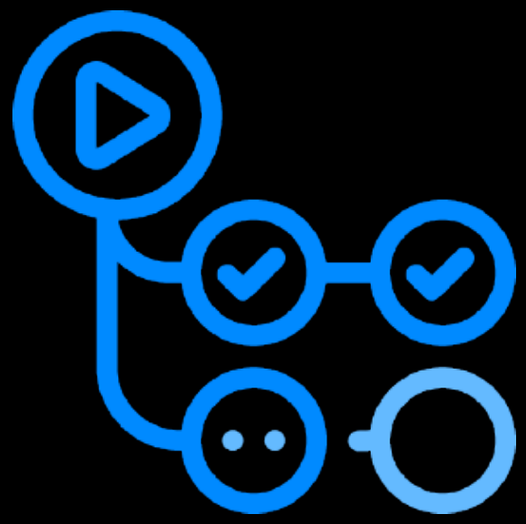


+



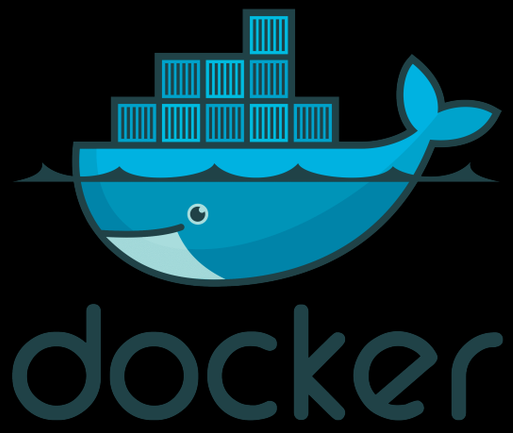
=

Actions + Docker



GitHub Actions

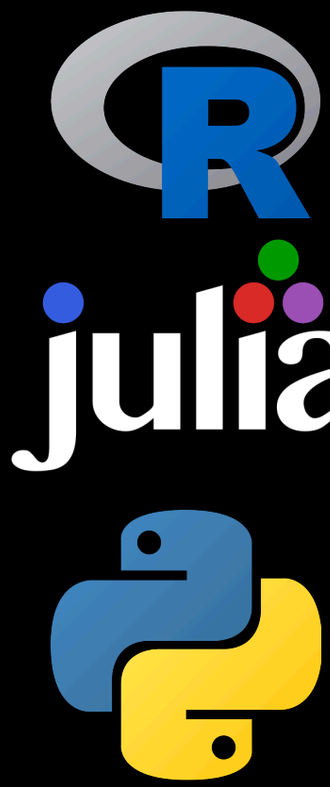
- Scheduler
- CI/CD

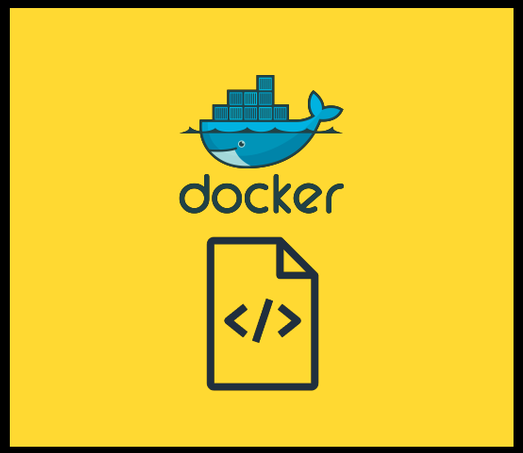


Docker

- Container
- CI/CD

Docker In A Nutshell

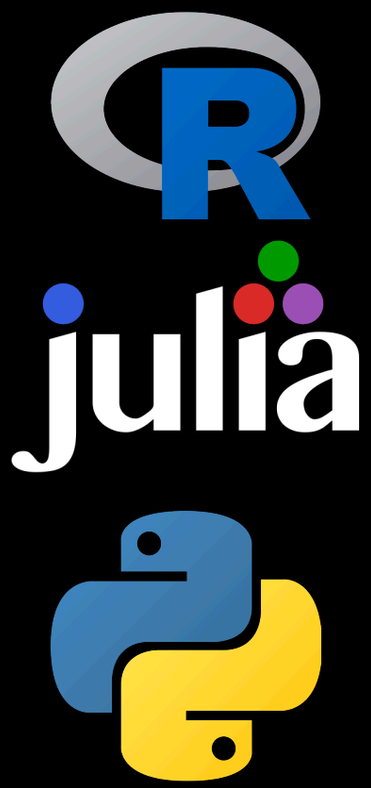


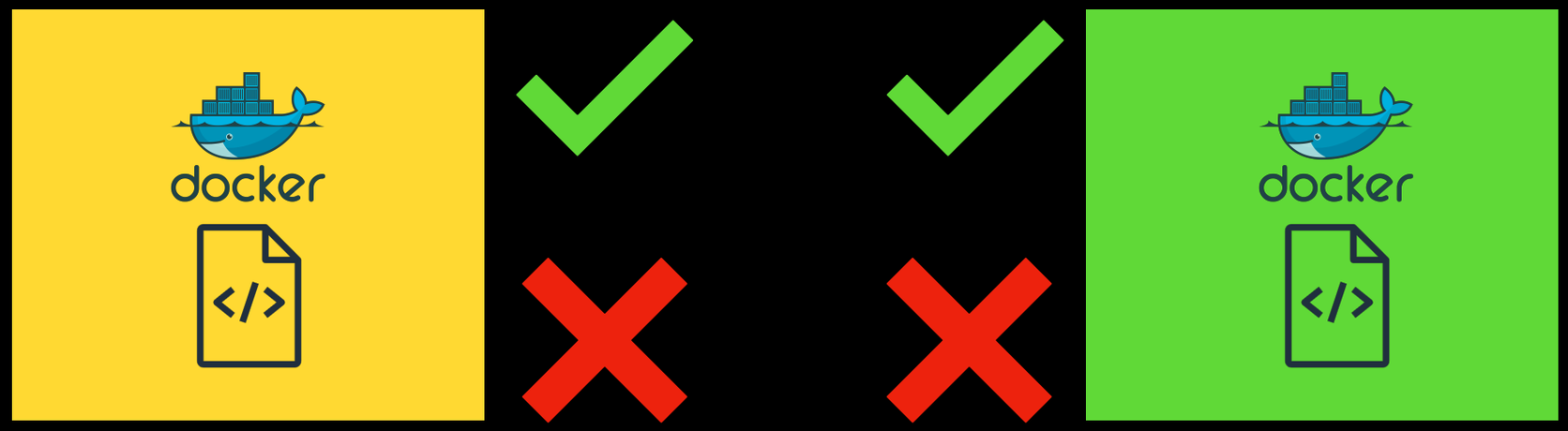


Local Env

Remote Env

Docker In A Nutshell



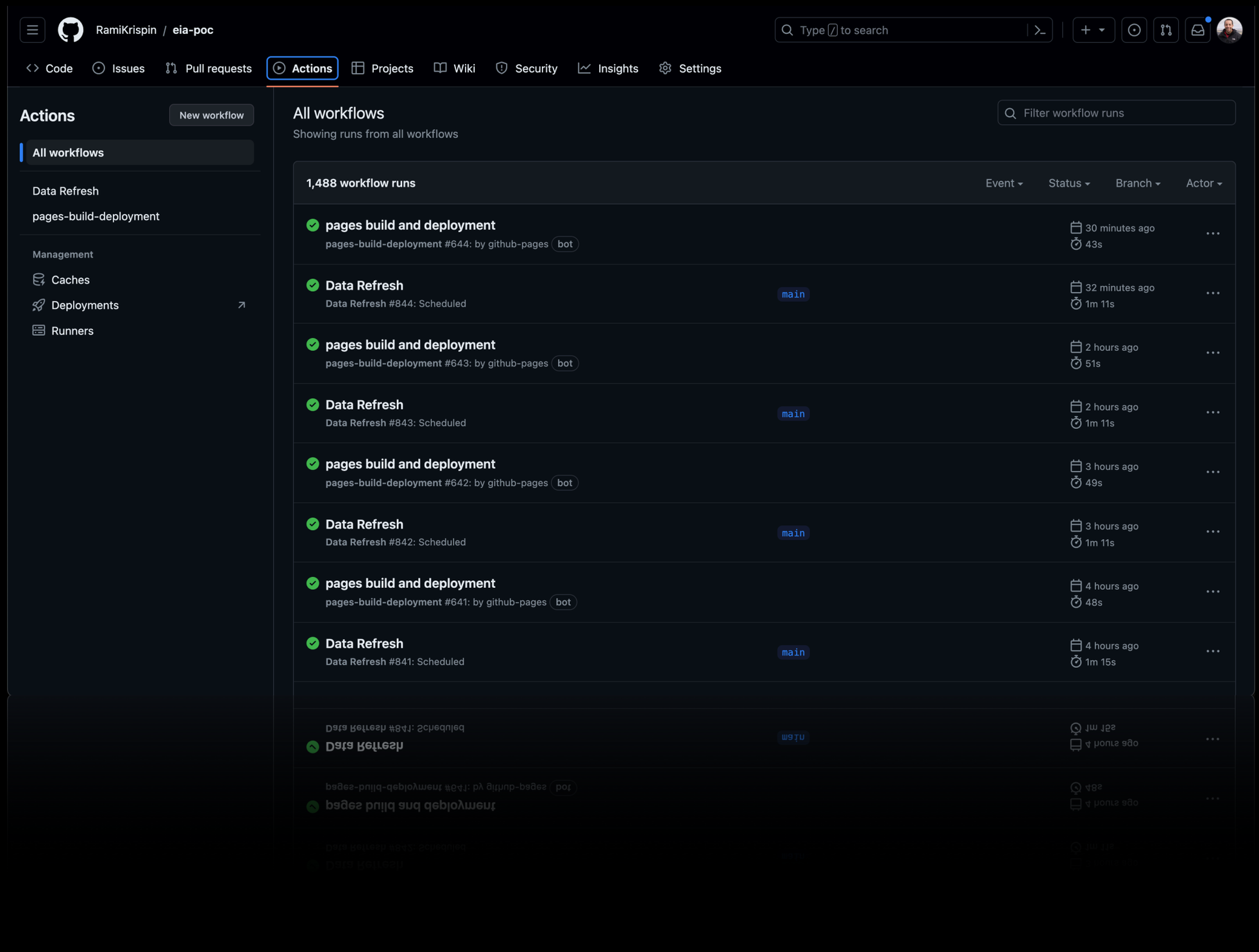


Local Env

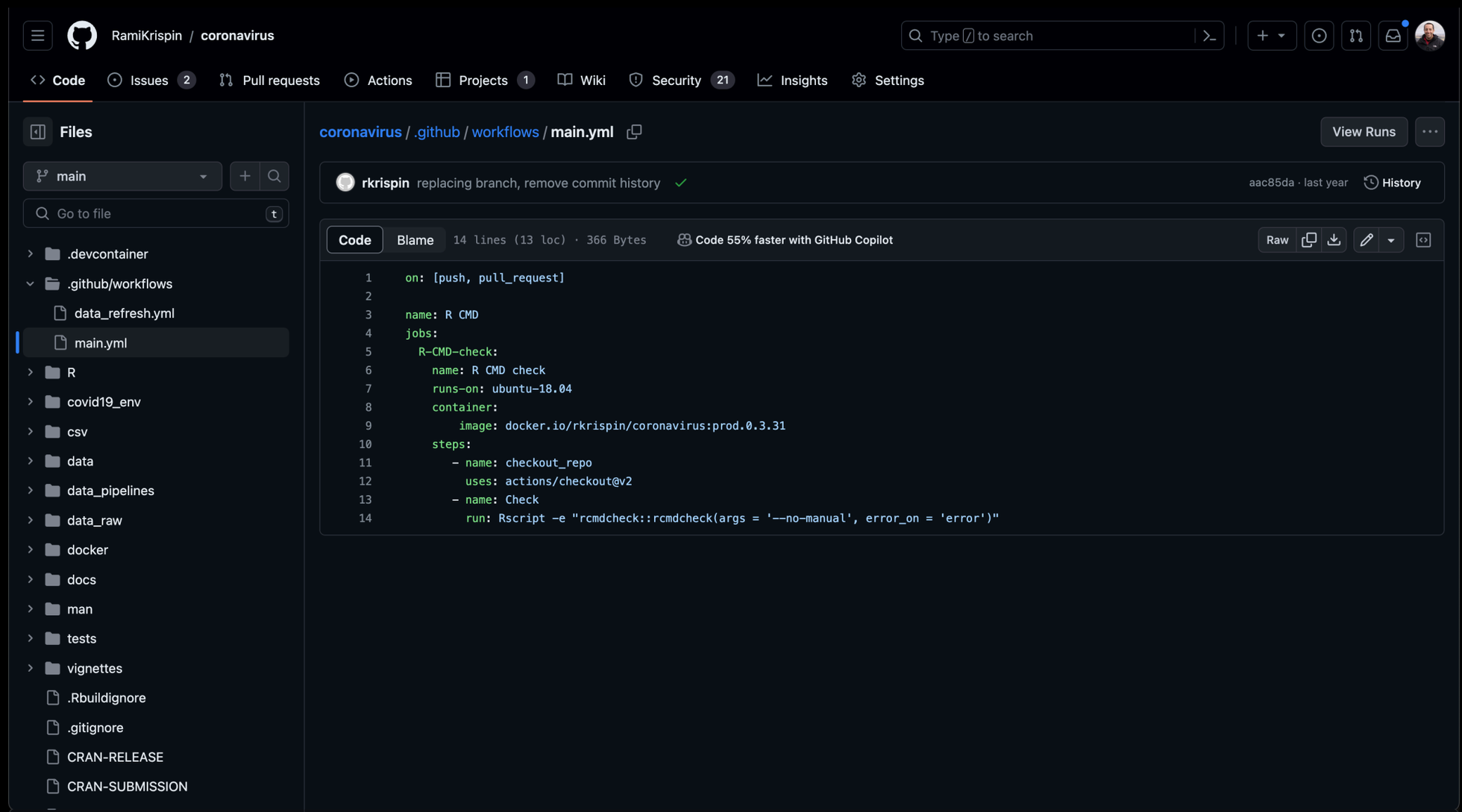
Remote Env

What Can You Build?

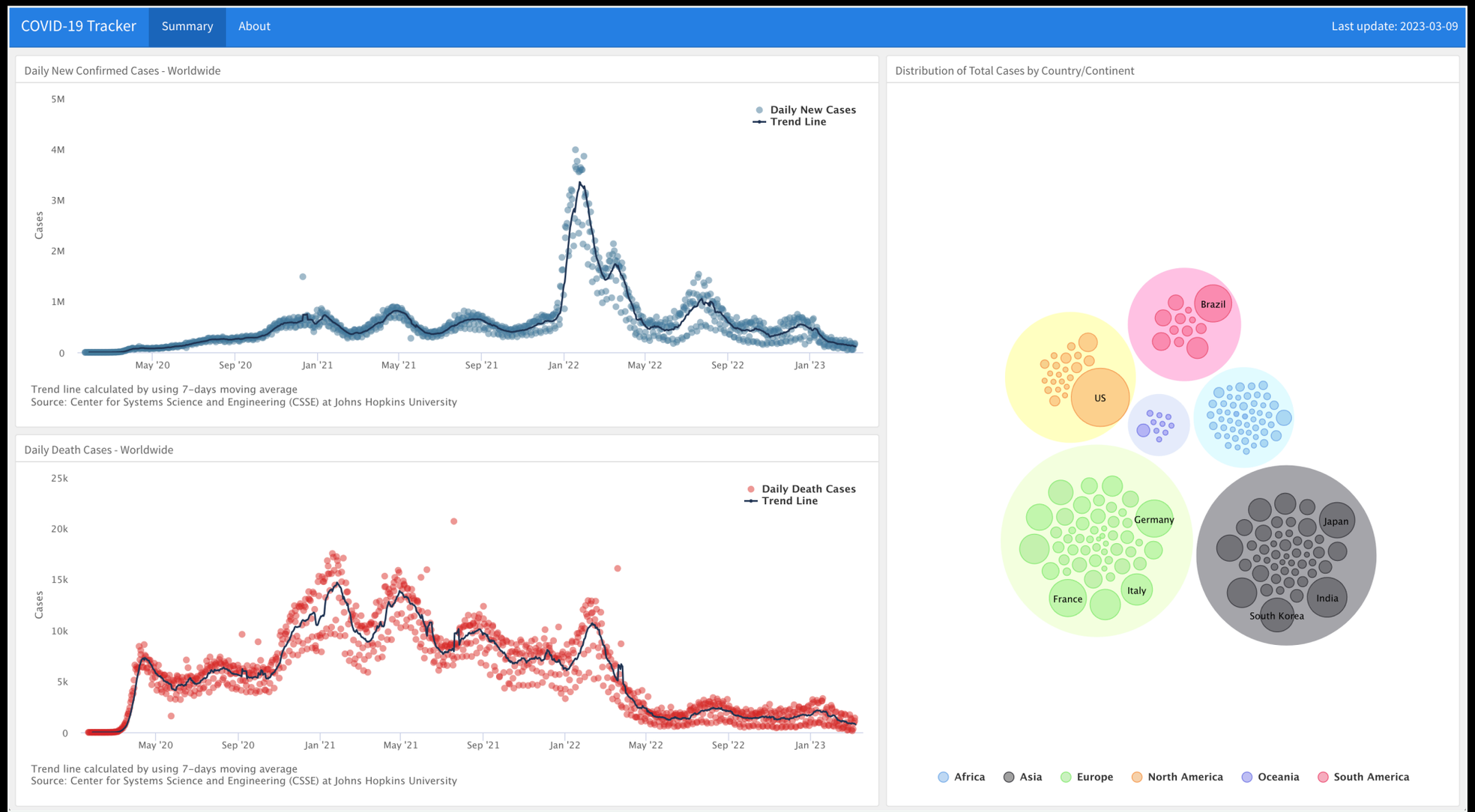
CI/CD & Automation



Code Development and Testing



Data Automation



ML Applications

Hugging Face 🤗

Select the Space SDK

You can choose between Streamlit, Gradio and Static for your Space. Or pick Docker to host any other app.

Streamlit

NEW

Gradio
3 templates

Docker
13 templates

Static
3 templates

Choose a Docker template:

Blank

JupyterLab

Argilla

Livebook

LabelStudio

AimStack

AutoTrain

Shiny (R)

Shiny (Python)

ZenML

ChatUI

Panel

Giskard

Quarto

Open and Free To Use

Limitations

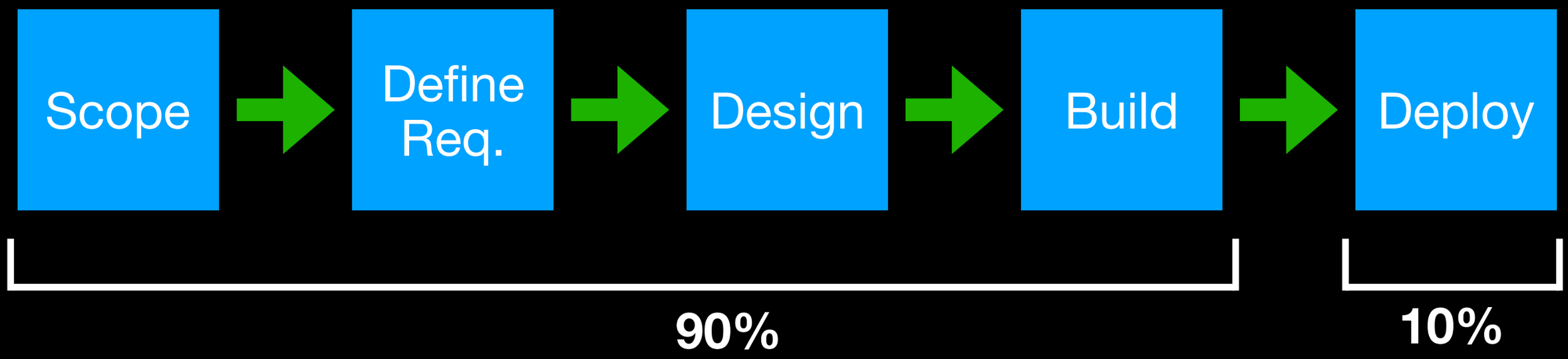
Limitations

- Not secure
- Compute power
- Availability
- Data storage

Deploy and Monitor ML and Data Pipelines with GitHub Actions

Building Pipeline - Workflow

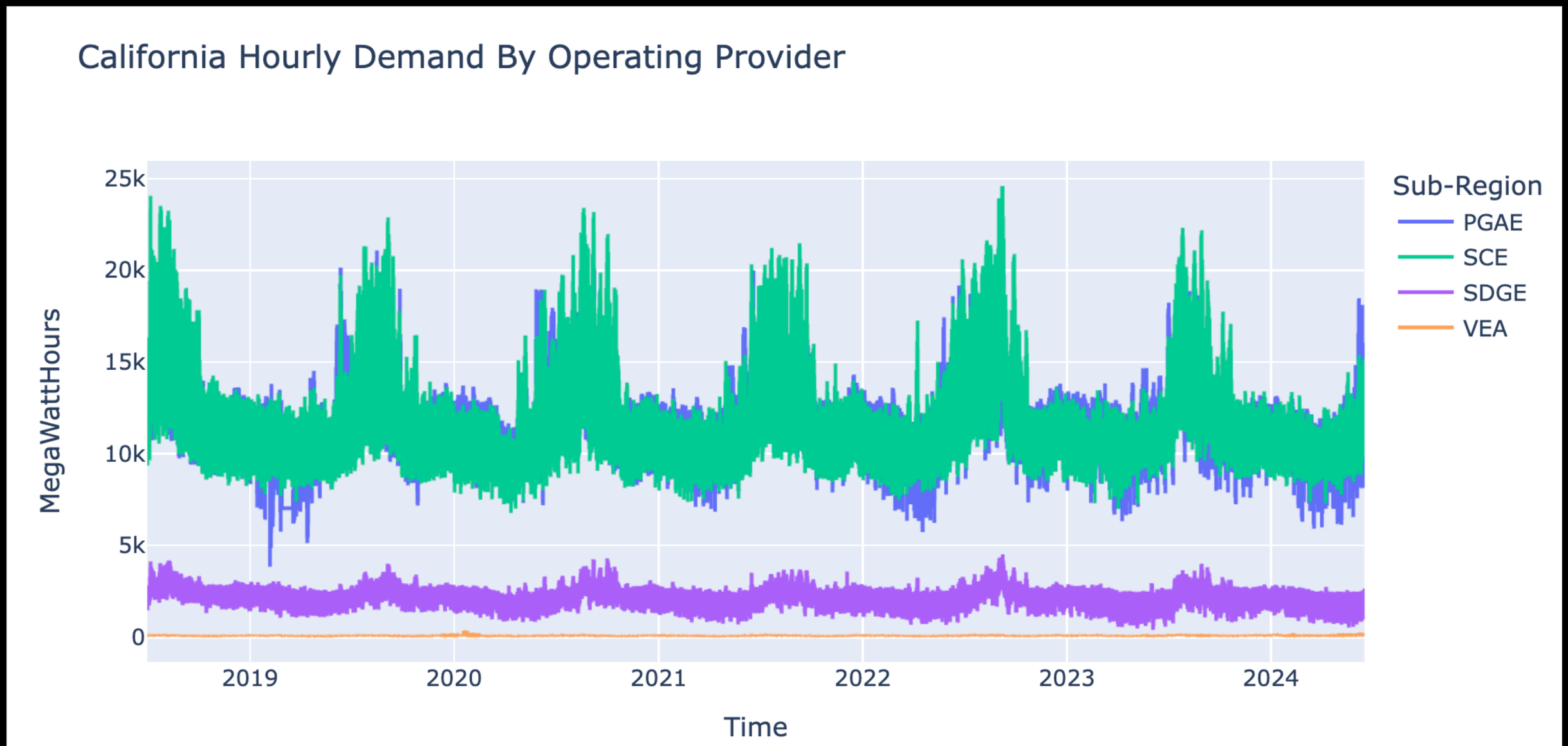
Building Pipeline - Workflow



Scope

- Forecast the hourly demand for electricity in California by sub-region
- This includes the following four providers:
    - Pacific Gas and Electric (PGAE)
    - Southern California Edison (SCE)
    - San Diego Gas and Electric (SDGE)
    - Valley Electric Association (VEA)
- Refresh the data every hour and the forecast every 24 hours

Scope


California Hourly Demand By Operating Provider
MegaWattHours
25k
20k
15k
10k
5k
0
2019
2020
2021
2022
2023
2024
Time
Sub-Region
PGAE
SCE
SDGE
VEA

Requirements

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

Data Source - EIA API

What's New

Refinery Capacity Report
Jun 14, 2024

Wholesale Electricity Market Data
Jun 13, 2024

Short-Term Energy Outlook
Jun 11, 2024

More ›

Coming Up

State Energy Data System (SEDS) Complete Annual Report

Annual Solar Photovoltaic Module

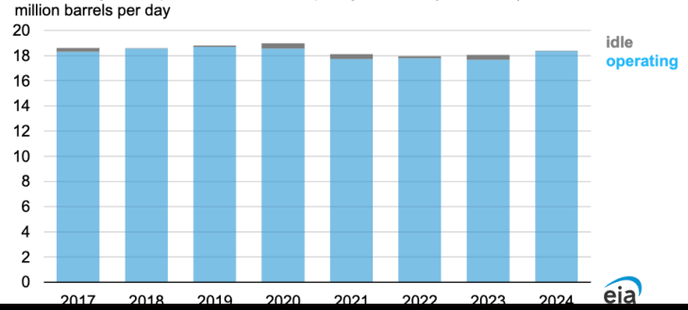
Today in Energy ›

Posted June 18, 2024

U.S. refinery capacity increased 2% in 2023 ›

Operable atmospheric crude oil distillation capacity, our primary measure of refinery capacity in the United States, increased by 2%, or 324,000 barrels per day (b/d), in 2023, according to our recently published *Refinery Capacity Report*. **More ›**

U.S. refinery atmospheric distillation capacity on Jan 1 (2017–2024)
million barrels per day



Data Highlights

WTI crude oil futures price
6/17/2024: **NA/barrel**
NA from week earlier
NA from year earlier

Natural gas futures price
6/17/2024: **NA/MMBtu**
NA from week earlier
NA from year earlier

Retail gasoline price
6/17/2024: **$3.435/gal**
▲ $0.006 from week earlier
▼ $0.142 from year earlier

Retail diesel price
6/17/2024: **$3.735/gal**
▲ $0.077 from week earlier
▼ $0.080 from year earlier

Weekly coal production

Data Source - EIA API

eia + Sources & Uses + Topics + Geography Search eia.gov

SUBMIT RESET

API URL:

```
https://api.eia.gov/v2/electricity/rto/region-sub-ba-data/data/?
frequency=hourly&data[0]=value&facets[parent][]=CISO&facets[subba]
[]=PGAE&facets[subba][]=SCE&facets[subba][]=SDGE&facets[subba]
[]=VEA&sort[0][column]=period&sort[0]
[direction]=desc&offset=0&length=5000
```

METHOD: GET

SERIES DESCRIPTION: Hourly demand by balancing authority subregion. Source: Form EIA-930 Product: Hourly Electric Grid Monitor

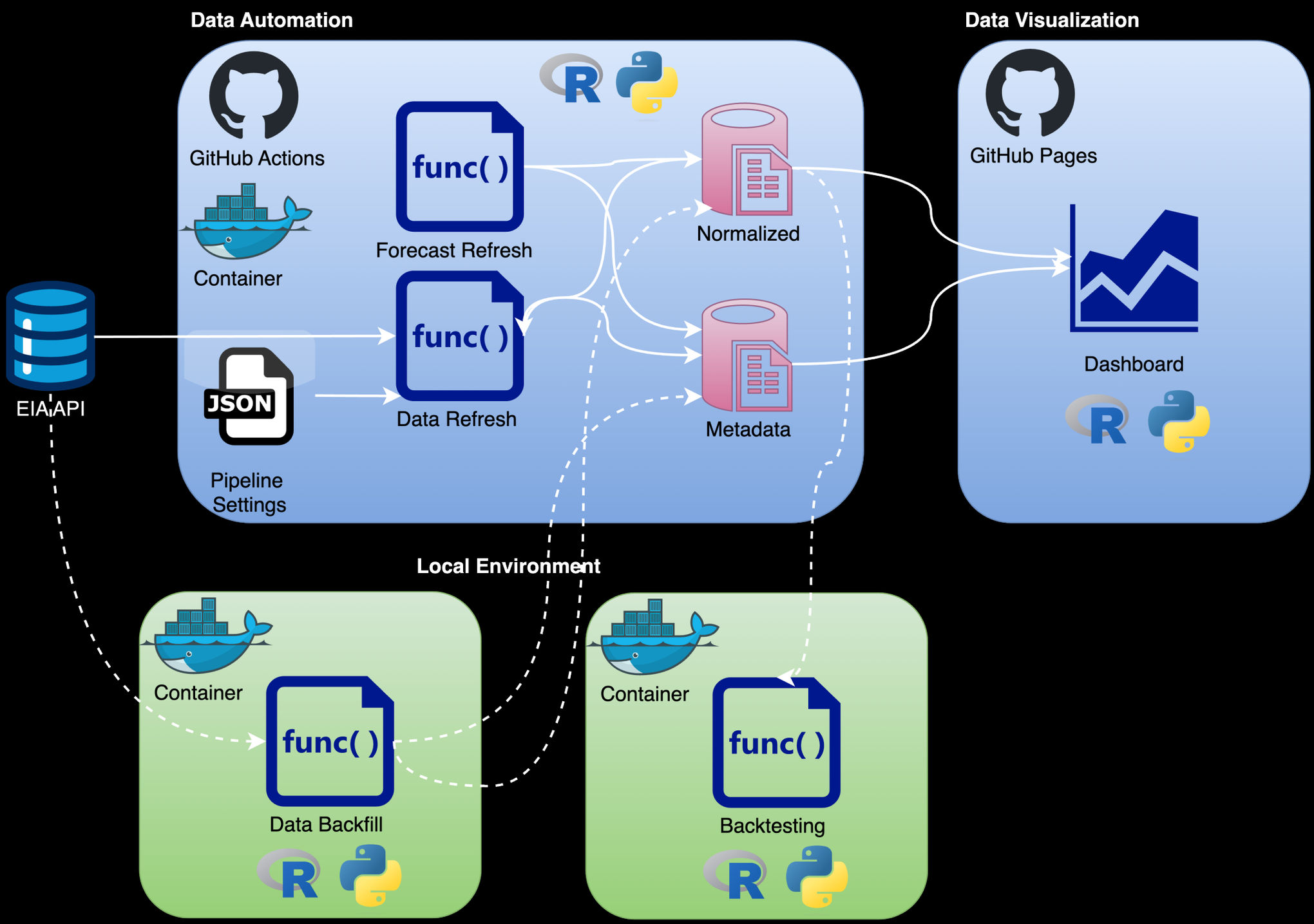
Need help using our API? Read our API Documentation to learn more.

HEADER:

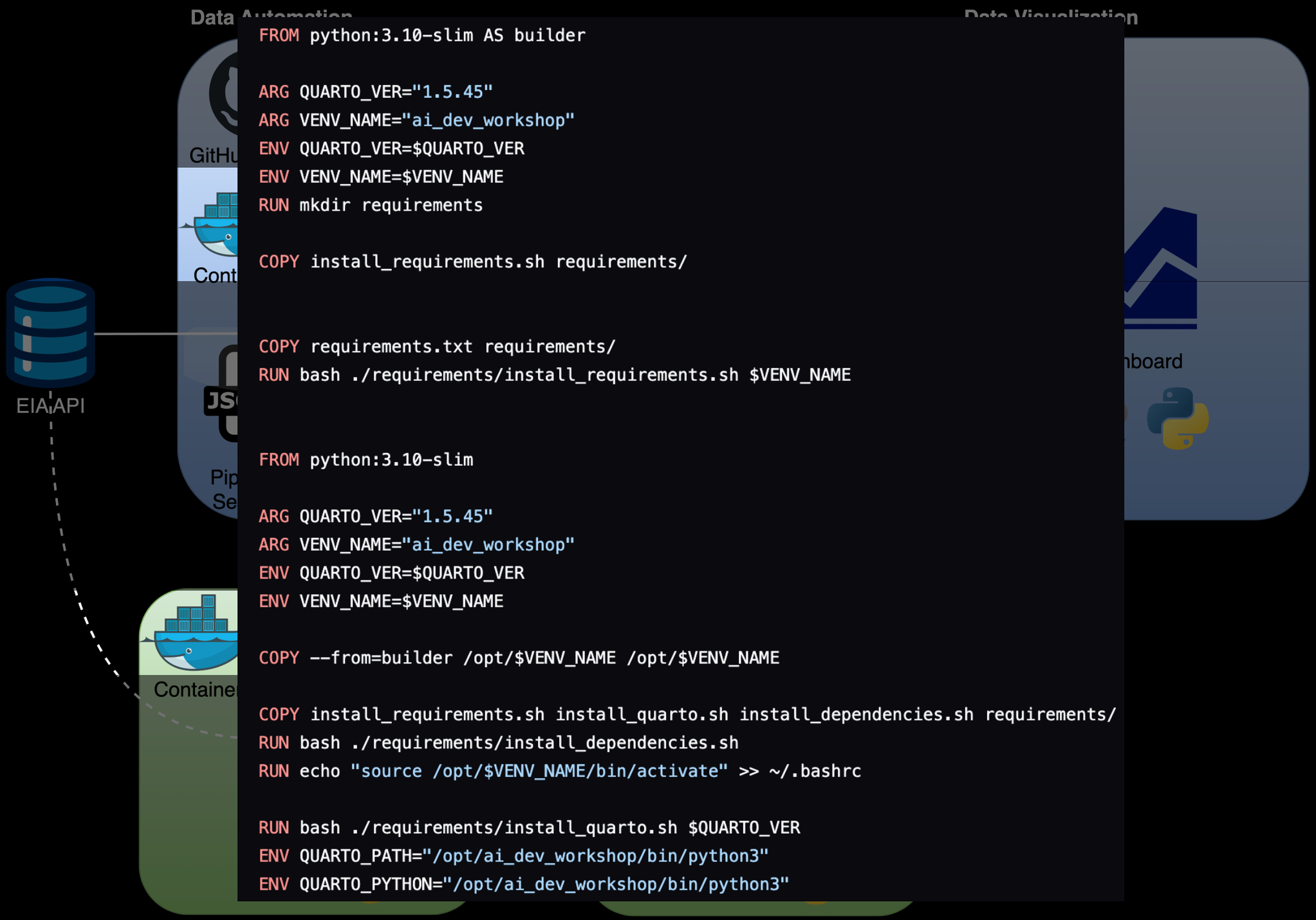
```
X-Params: {
    "frequency": "hourly",
    "data": [
        "value"
    ],
    "facets": {
        "parent": [
            "CISO"
        ],
        "subba": [
            "PGAE",
            "SCE",
            "SDGE",
            "VEA"
        ]
    },
    "start": null,
    "end": null,
    "sort": [
        {
            "column": "period",
            "direction": "desc"
        }
    ],
    "offset": 0,
    "length": 5000
}
```

Design

Pipeline Design

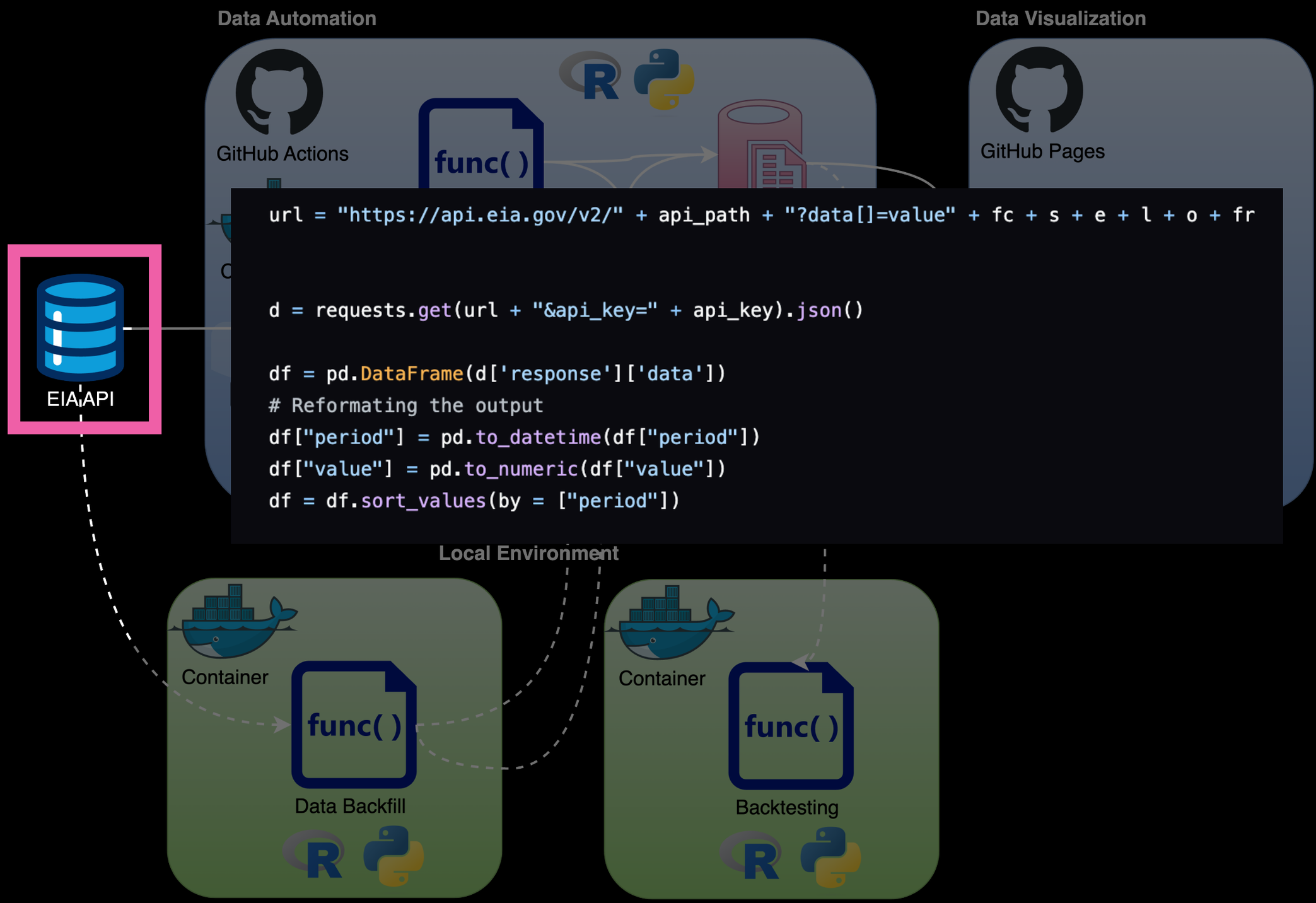


Pipeline Design



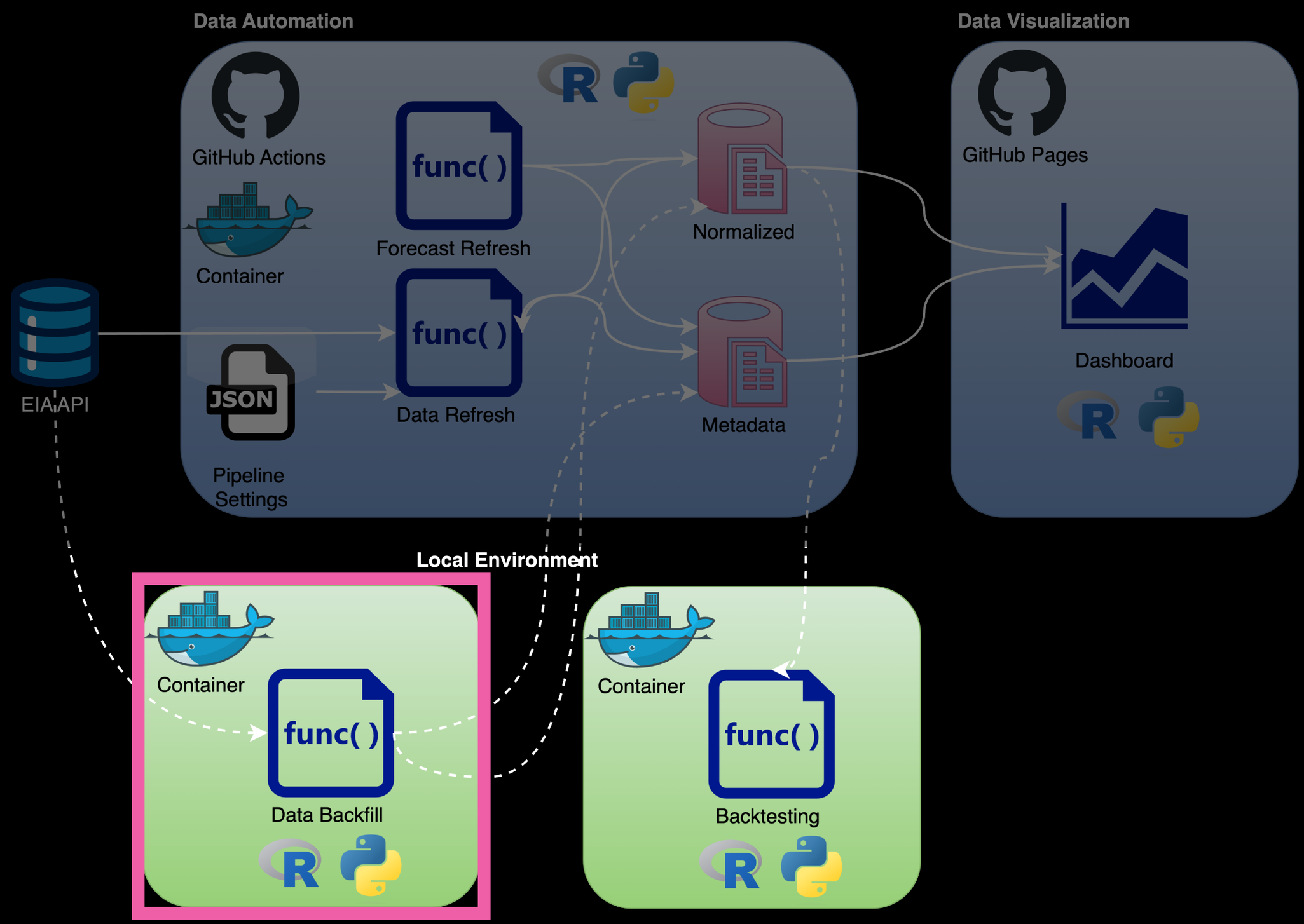

```dockerfile
FROM python:3.10-slim AS builder

ARG QUARTO_VER="1.5.45"
ARG VENV_NAME="ai_dev_workshop"
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="1.5.45"
ARG VENV_NAME="ai_dev_workshop"
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.sh install_dependencies.sh requirements/
RUN bash ./requirements/install_dependencies.sh
RUN echo "source /opt/$VENV_NAME/bin/activate" >> ~/.bashrc

RUN bash ./requirements/install_quarto.sh $QUARTO_VER
ENV QUARTO_PATH="/opt/ai_dev_workshop/bin/python3"
ENV QUARTO_PYTHON="/opt/ai_dev_workshop/bin/python3"
```

Pipeline Design

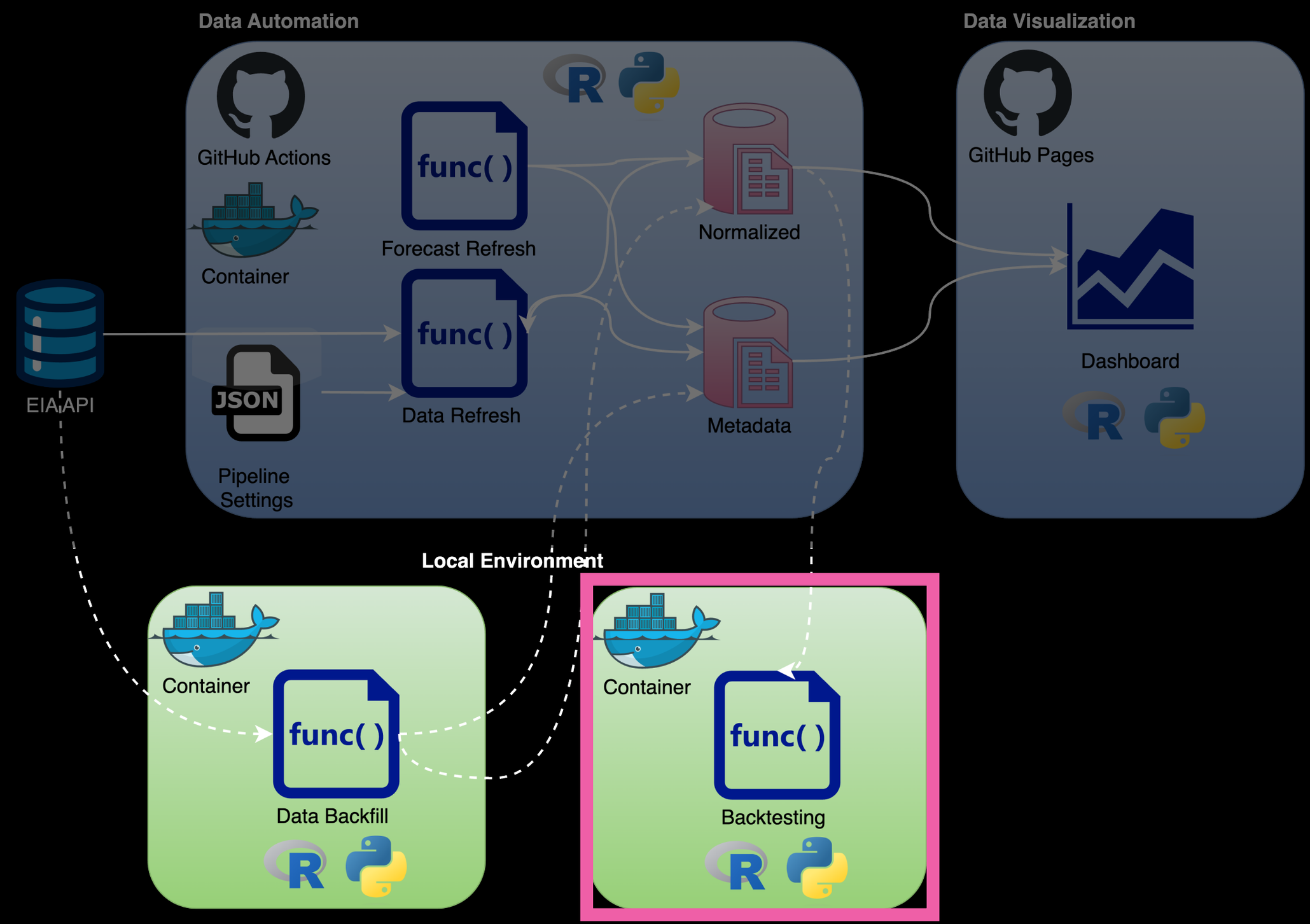



```
url = "https://api.eia.gov/v2/" + api_path + "?data[]=value" + fc + s + e + l + o + fr


d = requests.get(url + "&api_key=" + api_key).json()

df = pd.DataFrame(d['response']['data'])
# Reformating the output
df["period"] = pd.to_datetime(df["period"])
df["value"] = pd.to_numeric(df["value"])
df = df.sort_values(by = ["period"])
```

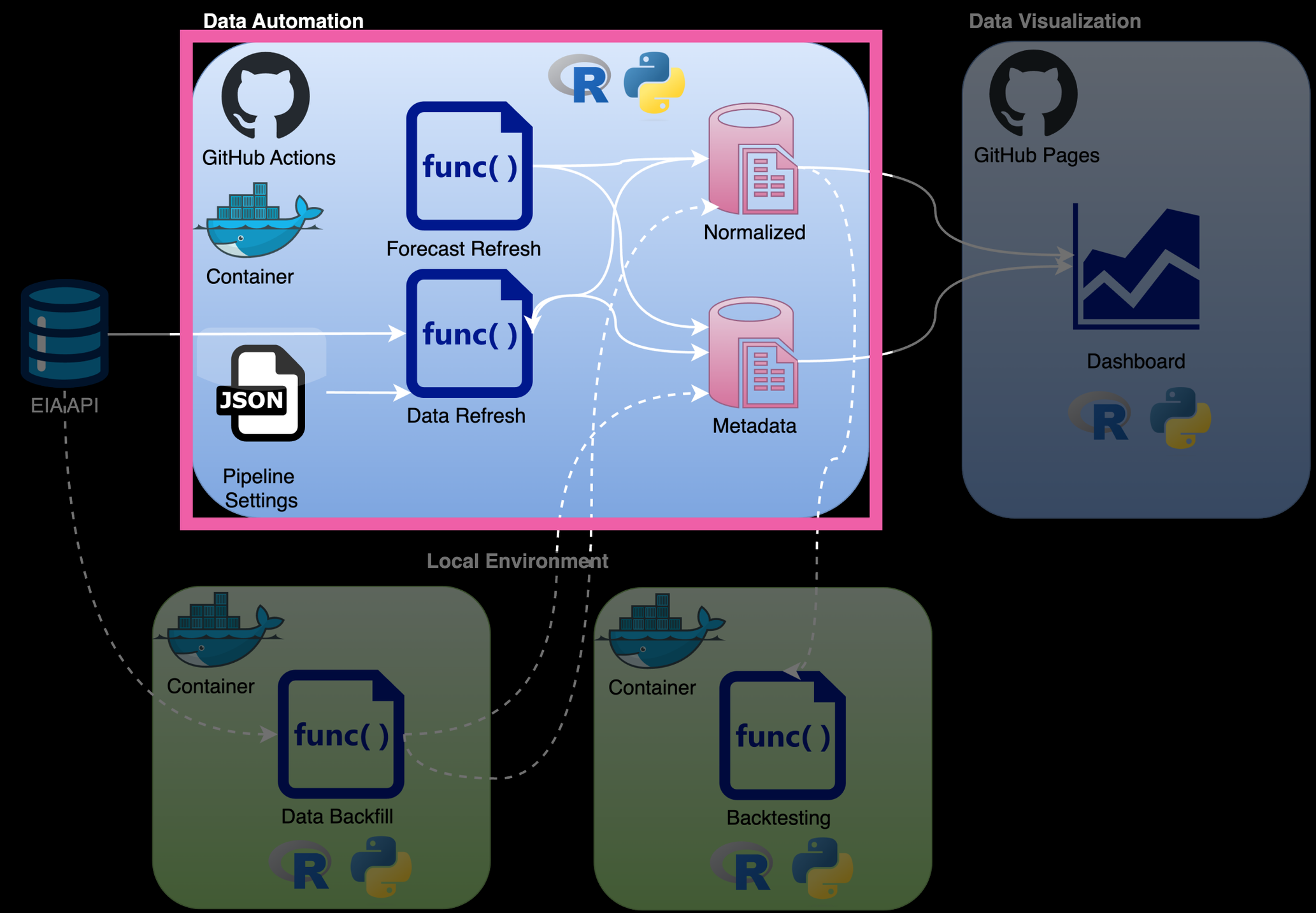
Pipeline Design



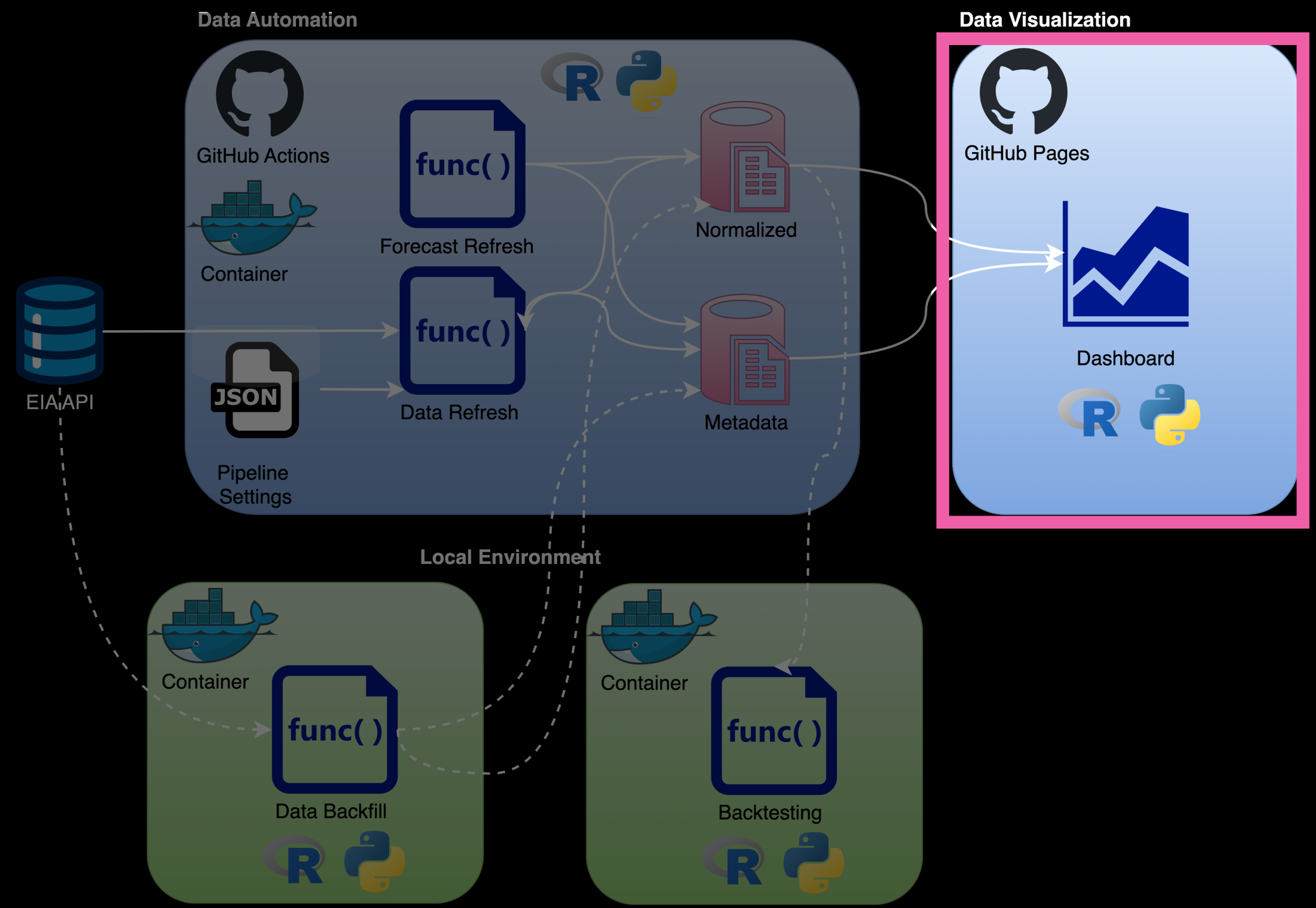
Pipeline Design



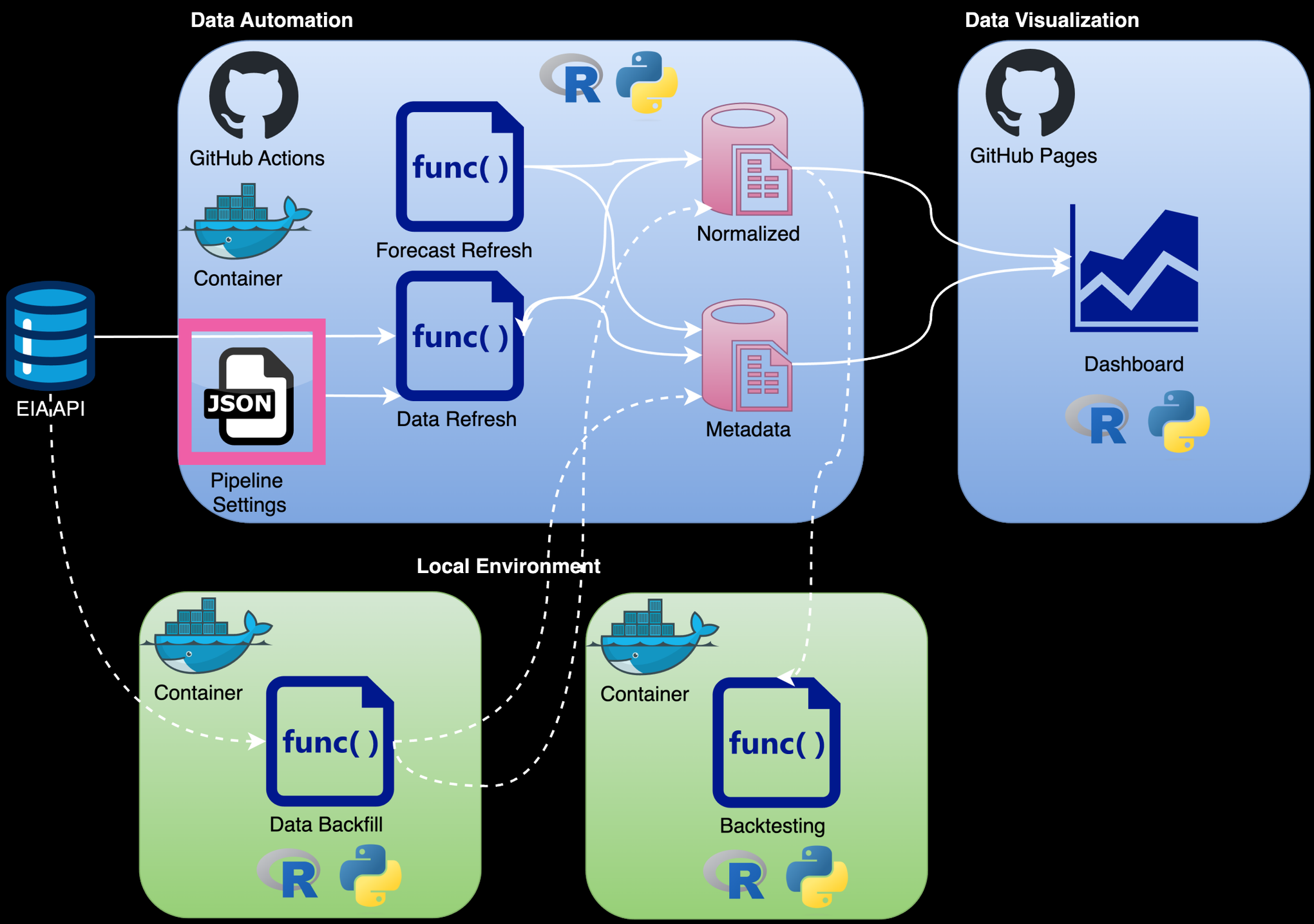
Pipeline Design



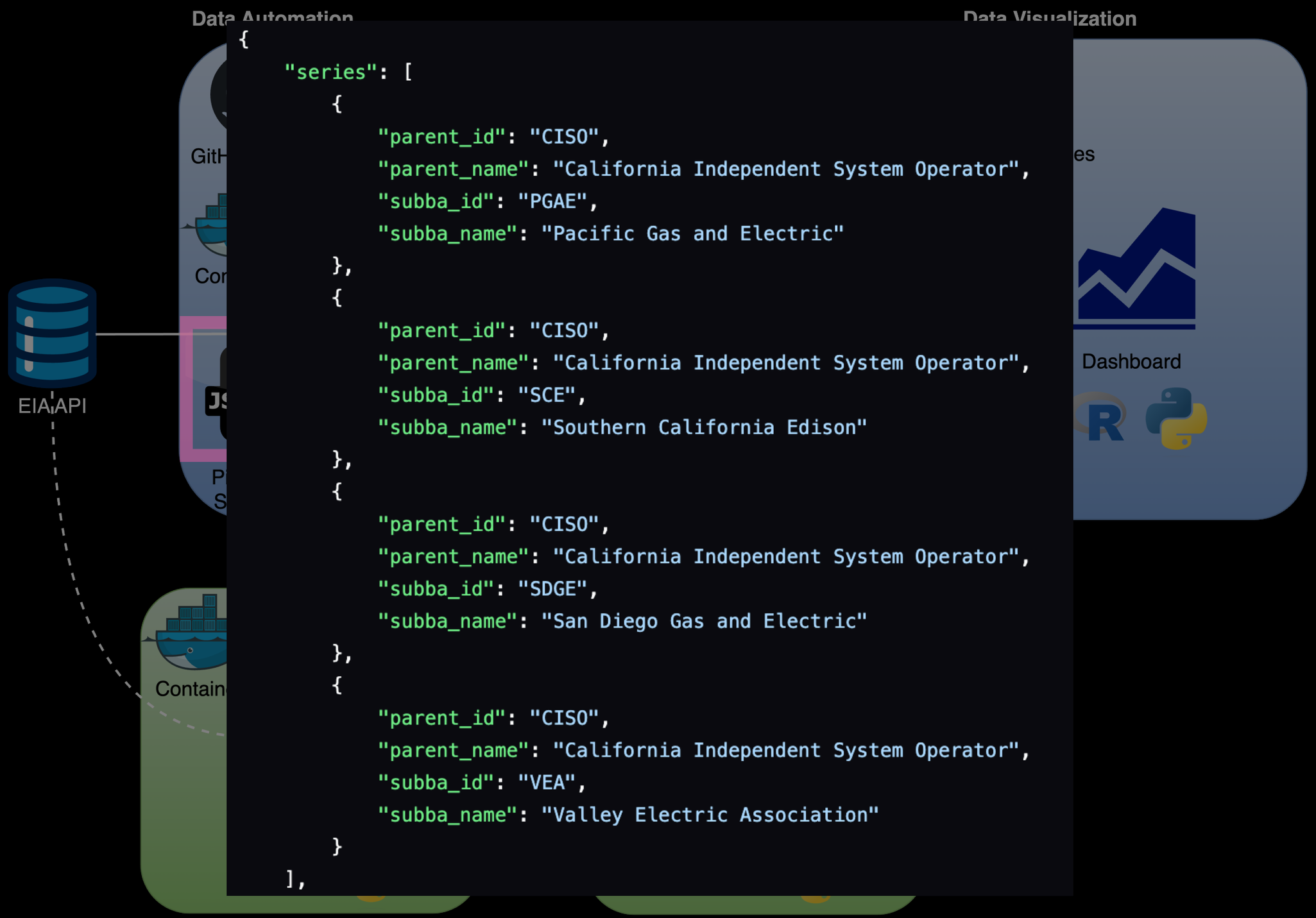
Pipeline Design



Pipeline Design



Pipeline Design

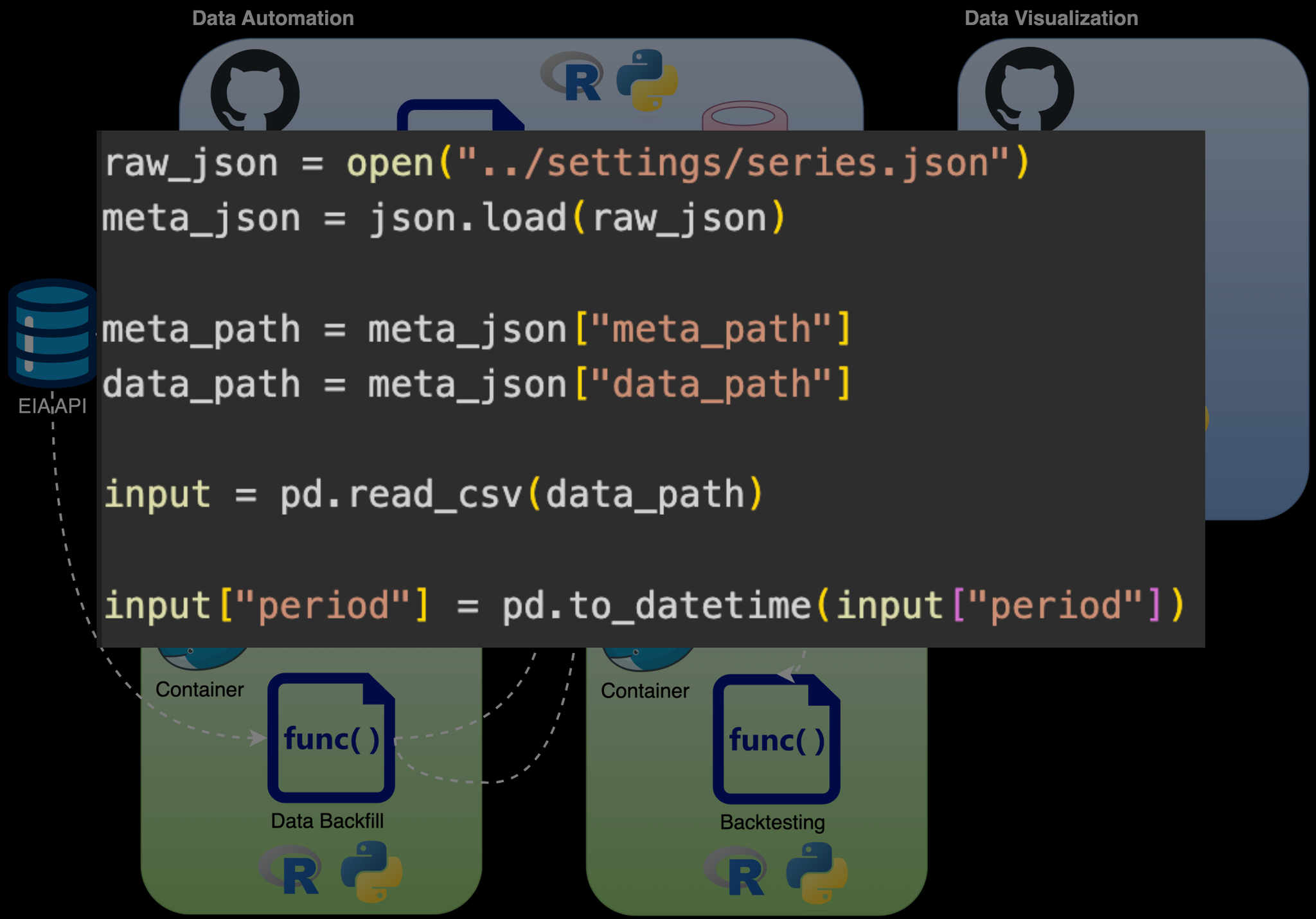



```
{
    "series": [
        {
            "parent_id": "CISO",
            "parent_name": "California Independent System Operator",
            "subba_id": "PGAE",
            "subba_name": "Pacific Gas and Electric"
        },
        {
            "parent_id": "CISO",
            "parent_name": "California Independent System Operator",
            "subba_id": "SCE",
            "subba_name": "Southern California Edison"
        },
        {
            "parent_id": "CISO",
            "parent_name": "California Independent System Operator",
            "subba_id": "SDGE",
            "subba_name": "San Diego Gas and Electric"
        },
        {
            "parent_id": "CISO",
            "parent_name": "California Independent System Operator",
            "subba_id": "VEA",
            "subba_name": "Valley Electric Association"
        }
    ],
```

Pipeline Design

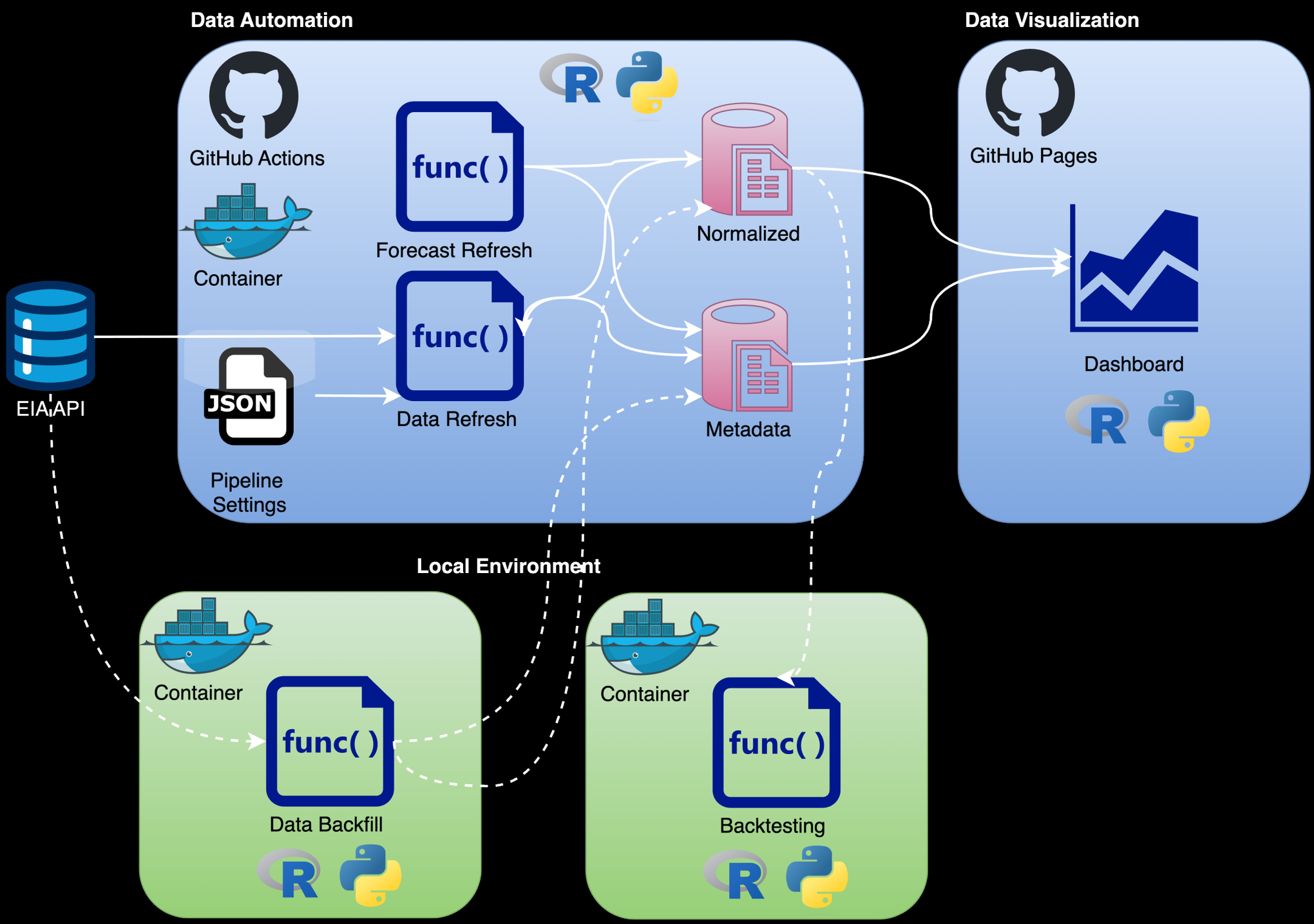
Pipeline Design

Pipeline Design

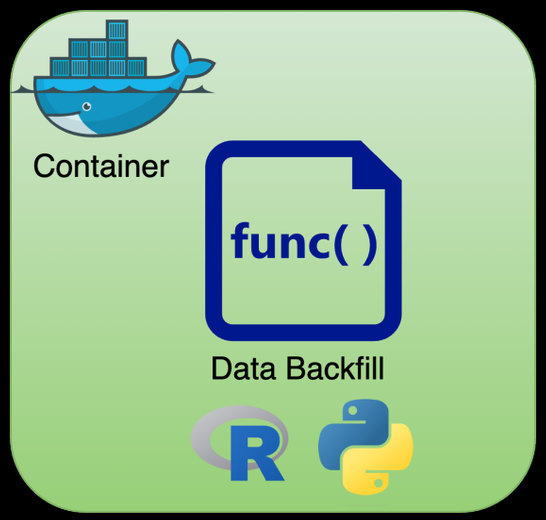



```
raw_json = open("../settings/series.json")
meta_json = json.load(raw_json)

meta_path = meta_json["meta_path"]
data_path = meta_json["data_path"]

input = pd.read_csv(data_path)

input["period"] = pd.to_datetime(input["period"])
```

Pipeline Design



Pipeline Design



- Initiate the pipeline
- Large dataset
- Use Quarto
- Data profiling

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

▼ Code

```
import eia_api as api
import eia_data
import pandas as pd
import numpy as np
import requests
import json
import os
import datetime
import plotly.express as px
```

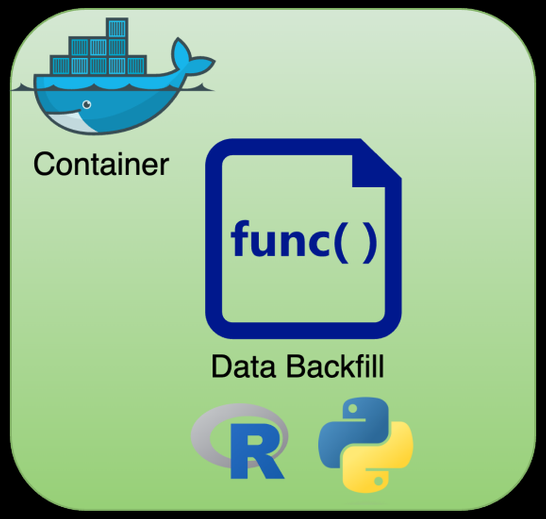
▼ Code

```
raw_json = open("../settings/series.json")
meta_json = json.load(raw_json)
series = pd.DataFrame(meta_json["series"])
api_path = meta_json["api_path"]
```

▼ Code

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

start = datetime.datetime(meta_json["start"]["year"],
 meta_json["start"]["month"],
 meta_json["start"]["day"],
 meta_json["start"]["hour"])
end = datetime.datetime(meta_json["end"]["year"],
```

Pipeline Design

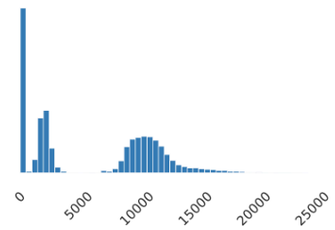


- Initiate the pipeline
- Large dataset
- Use Quarto
- Data profiling

value

Real number (ℝ)

Distinct	15375	**Minimum**	0
Distinct (%)	7.4%	**Maximum**	24595
Missing	392	**Zeros**	256
Missing (%)	0.2%	**Zeros (%)**	0.1%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6247.2254	**Memory size**	3.2 MiB



More details

Statistics | Histogram | Common values | Extreme values

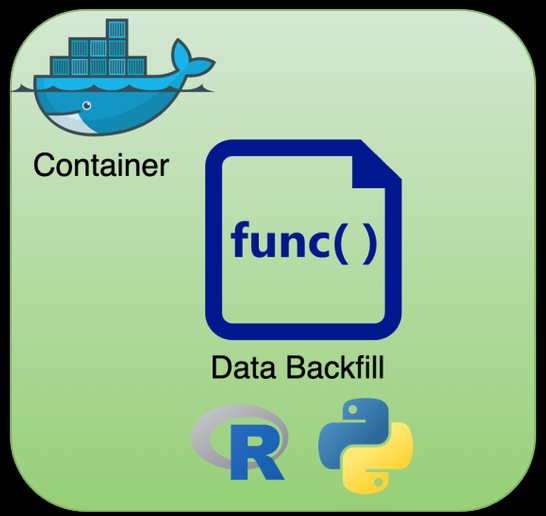
Quantile statistics

Minimum	0
5-th percentile	50
Q1	377.75
median	4486
Q3	10976
95-th percentile	14489.05
Maximum	24595
Range	24595
Interquartile range (IQR)	10598.25

Descriptive statistics

Standard deviation	5474.6113
Coefficient of variation (CV)	0.87632683
Kurtosis	-1.323918
Mean	6247.2254
Median Absolute Deviation (MAD)	4438
Skewness	0.25826644
Sum	1.3024216×10^9
Variance	29971368

Pipeline Design



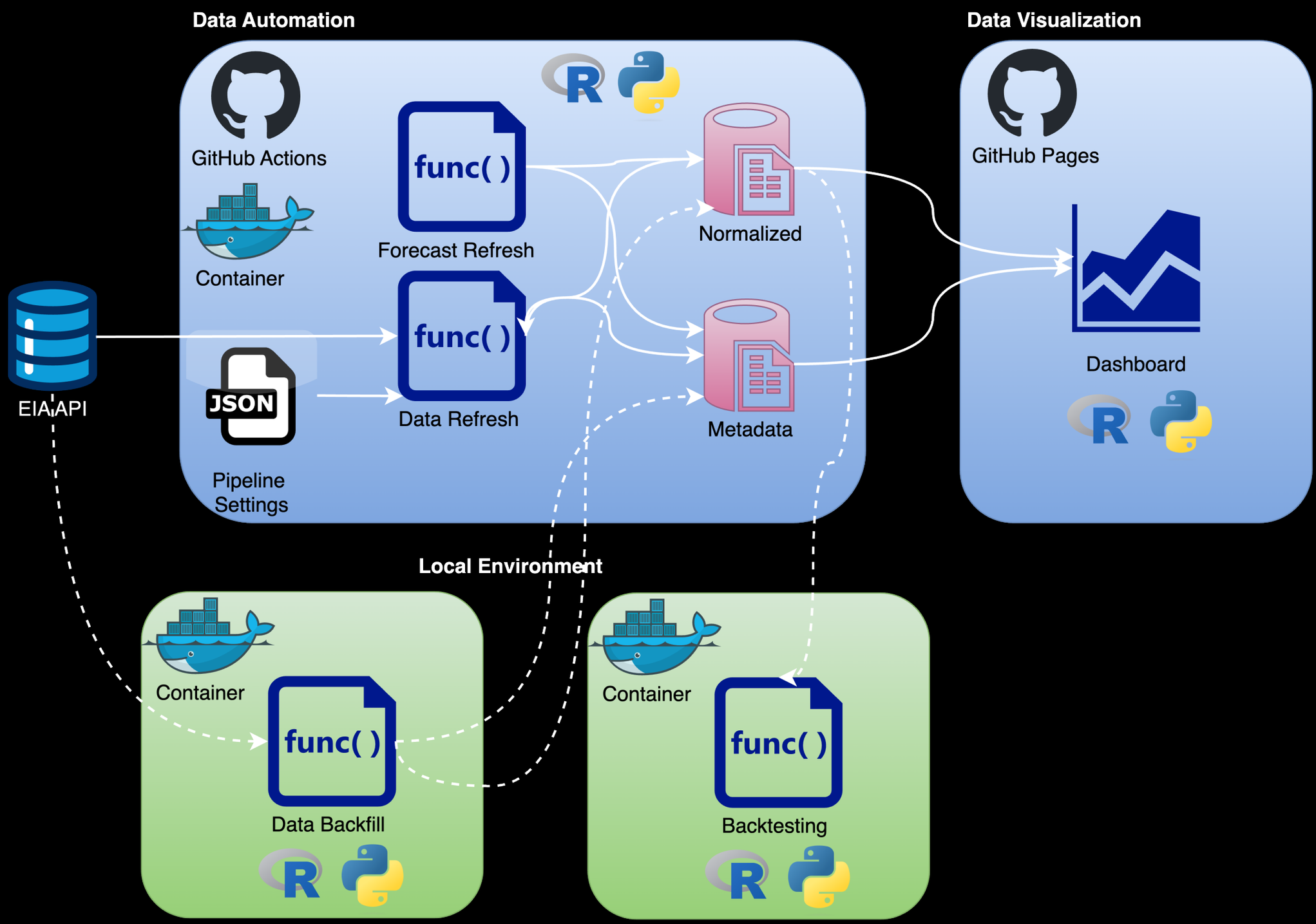
- Initiate the pipeline
- Large dataset
- Use Quarto
- Data profiling

functions > eia_api.py > ...

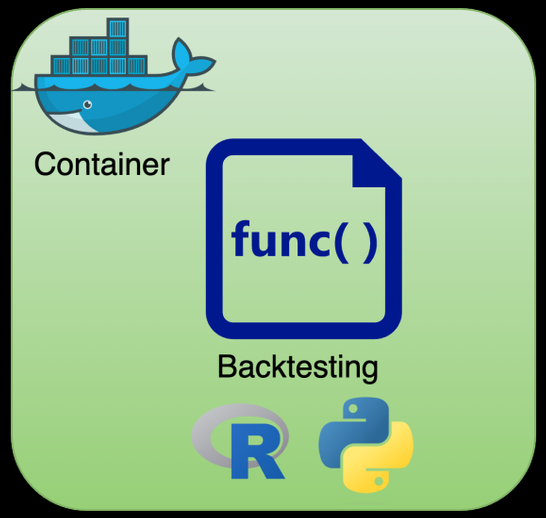
```python
import pandas as
import datetime
import requests

def day_offset(st
    current = [st
    while max(cur
        if(max(current) + datetime.timedelta(days= offset) < end):
            current.append(max(current) + datetime.timedelta(days= offset))
        else:
            current.append(end)

    return current

def hour_offset(start, end, offset):
    current = [start]
    while max(current) < end:
        if(max(current) + datetime.timedelta(hours = offset) < end):
            current.append(max(current) + datetime.timedelta(hours = offset
        else:
            current.append(end)

    return current

def eia_get(api_key,
            api_path,
            data = "value",
            facets = None,
            start = None,
            end = None,
            length = None,
            offset = None,
            frequency = None):
```

day_offset
hour_offset
eia_get
eia_backfill
eia_metadata

Pipeline Design



Pipeline Design



Model Backtesting

▶ Code

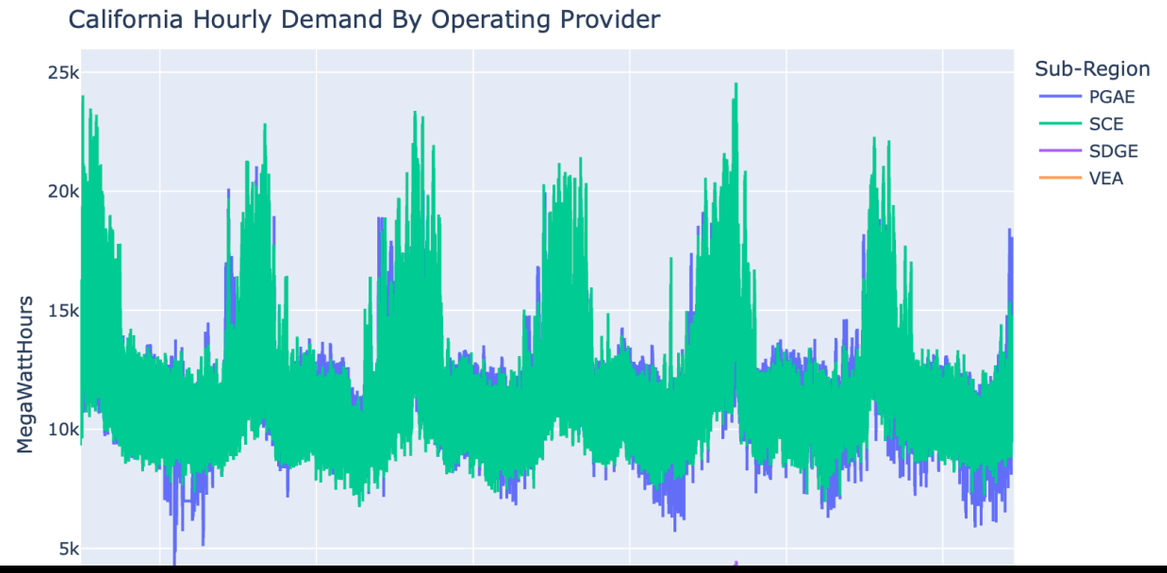
Load data

▶ Code

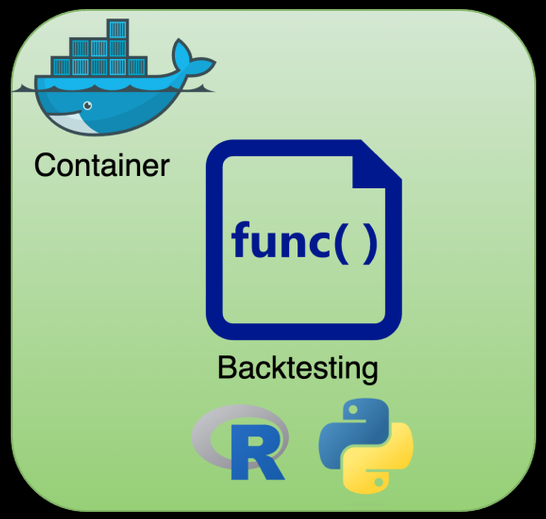
```
2024-06-14 23:00:00
```

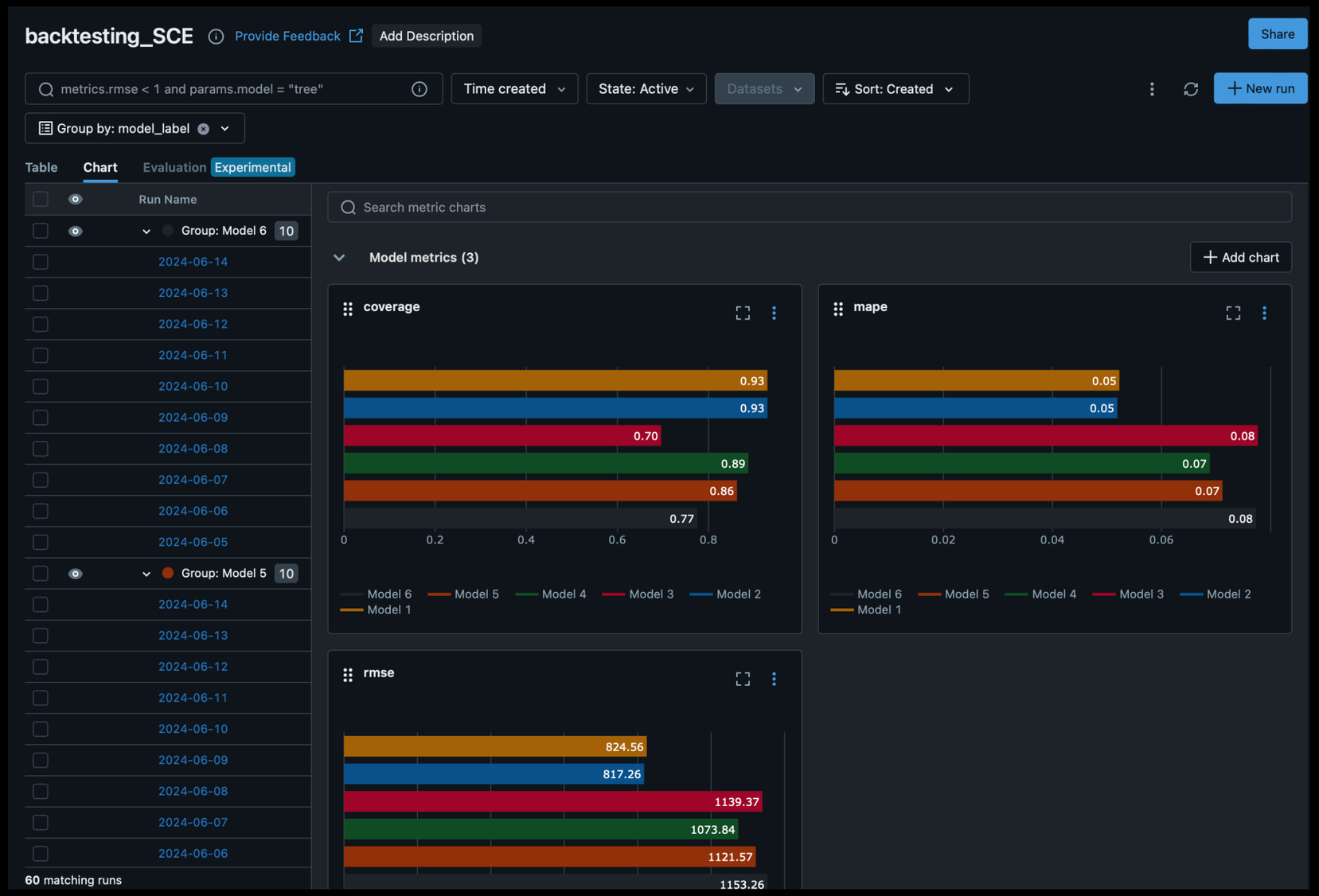
	period	subba	subba-name	parent	parent-name	value	value-units
0	2018-07-01 08:00:00	PGAE	Pacific Gas and Electric	CISO	California Independent System Operator	12522.0	megawatthours
1	2018-07-01 09:00:00	PGAE	Pacific Gas and Electric	CISO	California Independent System Operator	11745.0	megawatthours
2	2018-07-01 10:00:00	PGAE	Pacific Gas and Electric	CISO	California Independent System Operator	11200.0	megawatthours
3	2018-07-01 11:00:00	PGAE	Pacific Gas and Electric	CISO	California Independent System Operator	10822.0	megawatthours
4	2018-07-01 12:00:00	PGAE	Pacific Gas and Electric	CISO	California Independent System Operator	10644.0	megawatthours

▶ Code

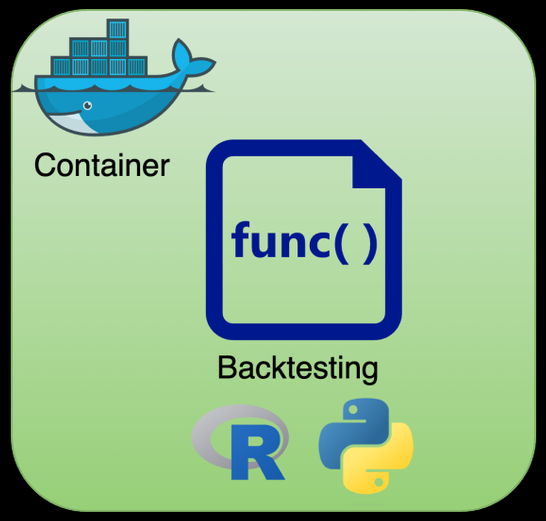


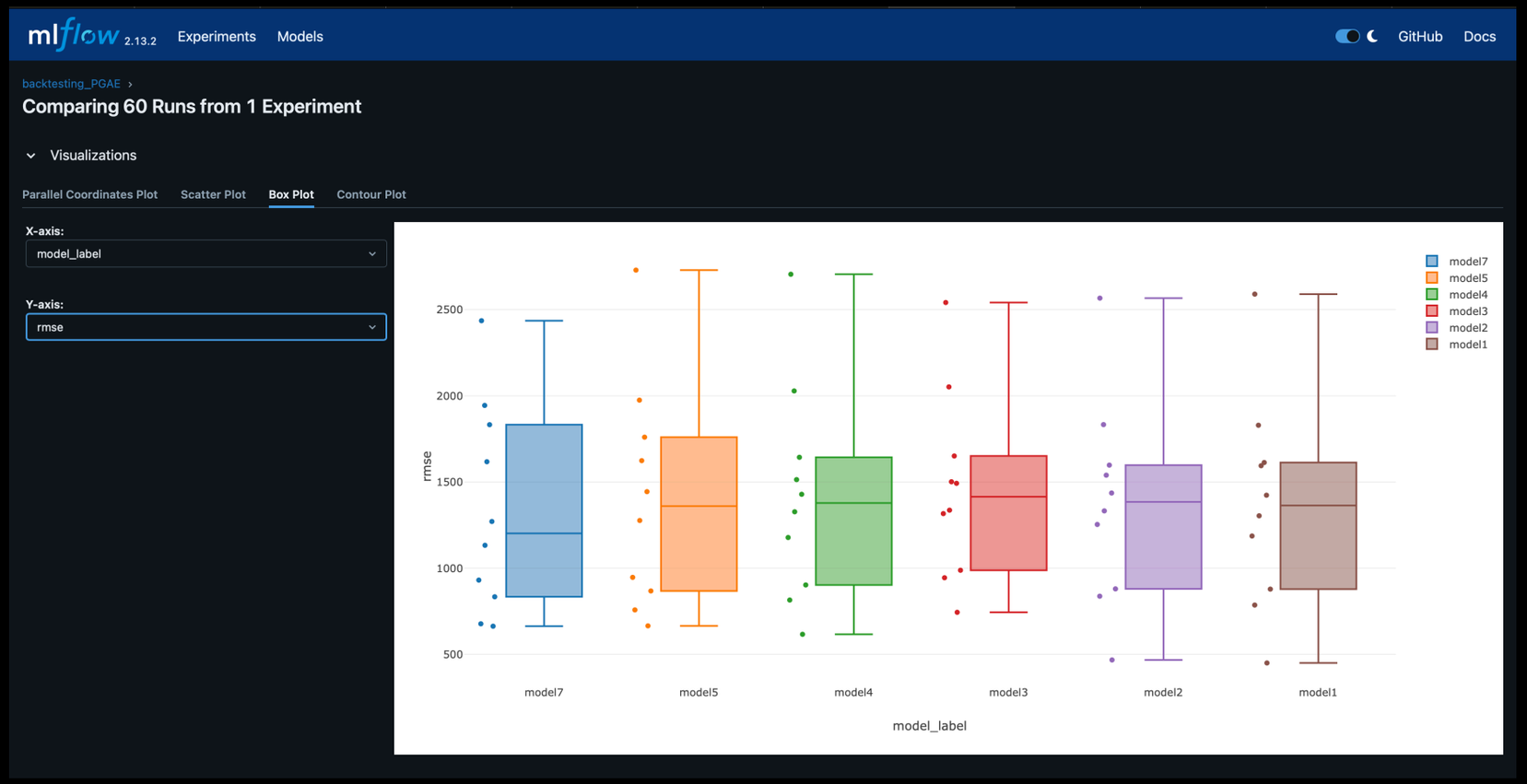
Pipeline Design



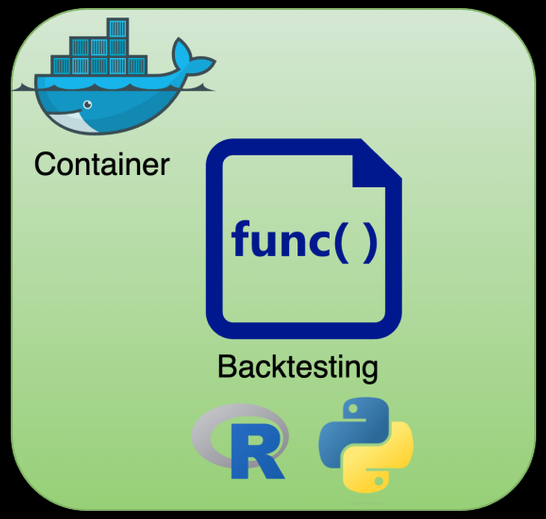


Pipeline Design



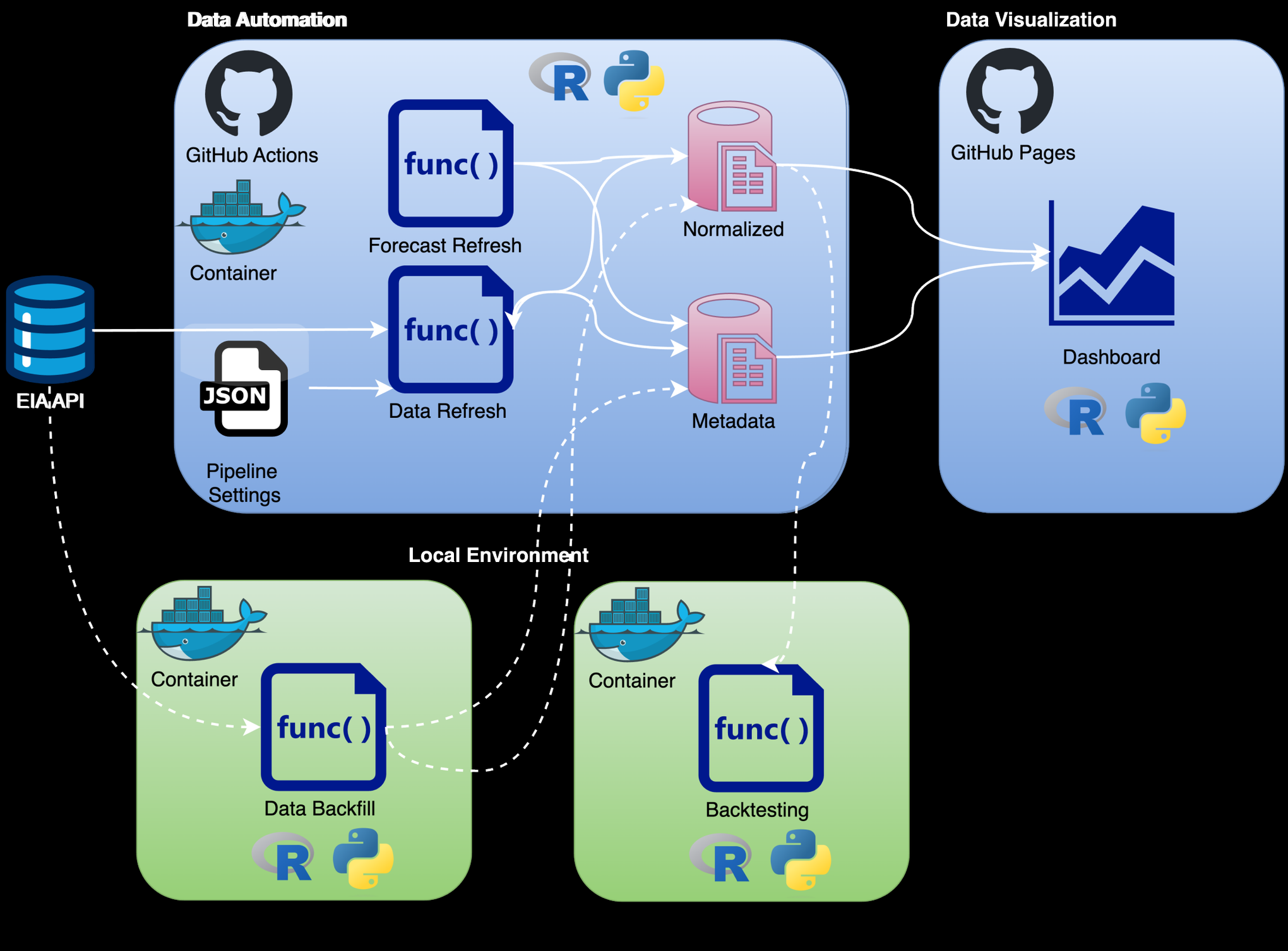


Pipeline Design



	forecast_label	model_label	model	mape	rmse	coverage	comments	subba
0	2024-06-14	model1	LinearRegressionModel	0.089334	1365.539586	0.725000	LM model with lags, training with 2 years of h...	PGAE
1	2024-06-14	model2	LinearRegressionModel	0.051878	817.262735	0.929167	LM model with lags, training with 3 years of h...	SCE
1	2024-06-14	model2	LinearRegressionModel	0.083141	188.657150	0.933333	LM model with lags, training with 3 years of h...	SDGE
0	2024-06-14	model1	LinearRegressionModel	0.100615	18.513913	0.887500	LM model with lags, training with 2 years of h...	VEA

Pipeline Design



Pipeline Design

RamiKrispin Auto update of the data ✓ 14f259e · 22 minutes ago History

Preview | Code | Blame 361 lines (361 loc) · 63.3 KB Code 55% faster with GitHub Copilot Raw

Search this file

respondent	respondent_type	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	type	update	success
US48	Demand	2024-06-04 09:09:29.847003	2015-07-01 05:00:00	2024-06-03 01:00:00	2015-07-01 05:00:00	2024-06-03 01:00:00	True	True	78237.0	0.0	backfill	True	True
US48	Demand	2024-06-04 09:13:35.486793	2024-06-03 02:00:00	2024-06-04 00:00:00	2024-05-27 02:00:00	2024-06-04 00:00:00	True	True	191.0	0.0	refresh	True	True
US48	Demand	2024-06-04 10:11:44.473285	2024-06-04 01:00:00	2024-06-04 01:00:00							refresh	False	False
US48	Demand	2024-06-04 11:10:51.547354	2024-06-04 01:00:00	2024-06-04 02:00:00	2024-05-28 01:00:00	2024-06-04 02:00:00	True	True	170.0	0.0	refresh	True	True
US48	Demand	2024-06-04 12:20:08.853848	2024-06-04 03:00:00	2024-06-04 03:00:00							refresh	False	False
US48	Demand	2024-06-04 13:13:27.580243	2024-06-04 03:00:00	2024-06-04 03:00:00							refresh	False	False
US48	Demand	2024-06-04 14:11:05.642034	2024-06-04 03:00:00	2024-06-04 23:00:00	2024-05-28 03:00:00	2024-06-04 11:00:00	True	False	177.0	0.0	refresh	True	True
US48	Demand	2024-06-04 15:12:08.108327	2024-06-04 12:00:00	2024-06-04 23:00:00	2024-05-28 12:00:00	2024-06-04 14:00:00	True	False	171.0	0.0	refresh	True	True
US48	Demand	2024-06-04 16:14:20.455581	2024-06-04 15:00:00	2024-06-04 23:00:00	2024-05-28 15:00:00	2024-06-04 15:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 17:10:30.755251	2024-06-04 16:00:00	2024-06-04 23:00:00	2024-05-28 16:00:00	2024-06-04 16:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 18:14:45.830771	2024-06-04 17:00:00	2024-06-04 23:00:00	2024-05-28 17:00:00	2024-06-04 17:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 19:09:33.113685	2024-06-04 18:00:00	2024-06-04 23:00:00	2024-05-28 18:00:00	2024-06-04 18:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 20:12:00.866350	2024-06-04 19:00:00	2024-06-04 23:00:00	2024-05-28 19:00:00	2024-06-04 19:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 21:10:53.584028	2024-06-04 20:00:00	2024-06-04 23:00:00	2024-05-28 20:00:00	2024-06-04 20:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 22:11:36.357769	2024-06-04 21:00:00	2024-06-04 23:00:00	2024-05-28 21:00:00	2024-06-04 21:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 23:12:31.360776	2024-06-04 22:00:00	2024-06-04 23:00:00	2024-05-28 22:00:00	2024-06-04 22:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-05 00:48:57.513133	2024-06-04 23:00:00	2024-06-04 23:00:00							refresh	False	False

Pipeline Design

What?

- Logs
- Unit tests

RamiKrispin Auto update of the data ✓ 14f259e · 22 minutes ago History

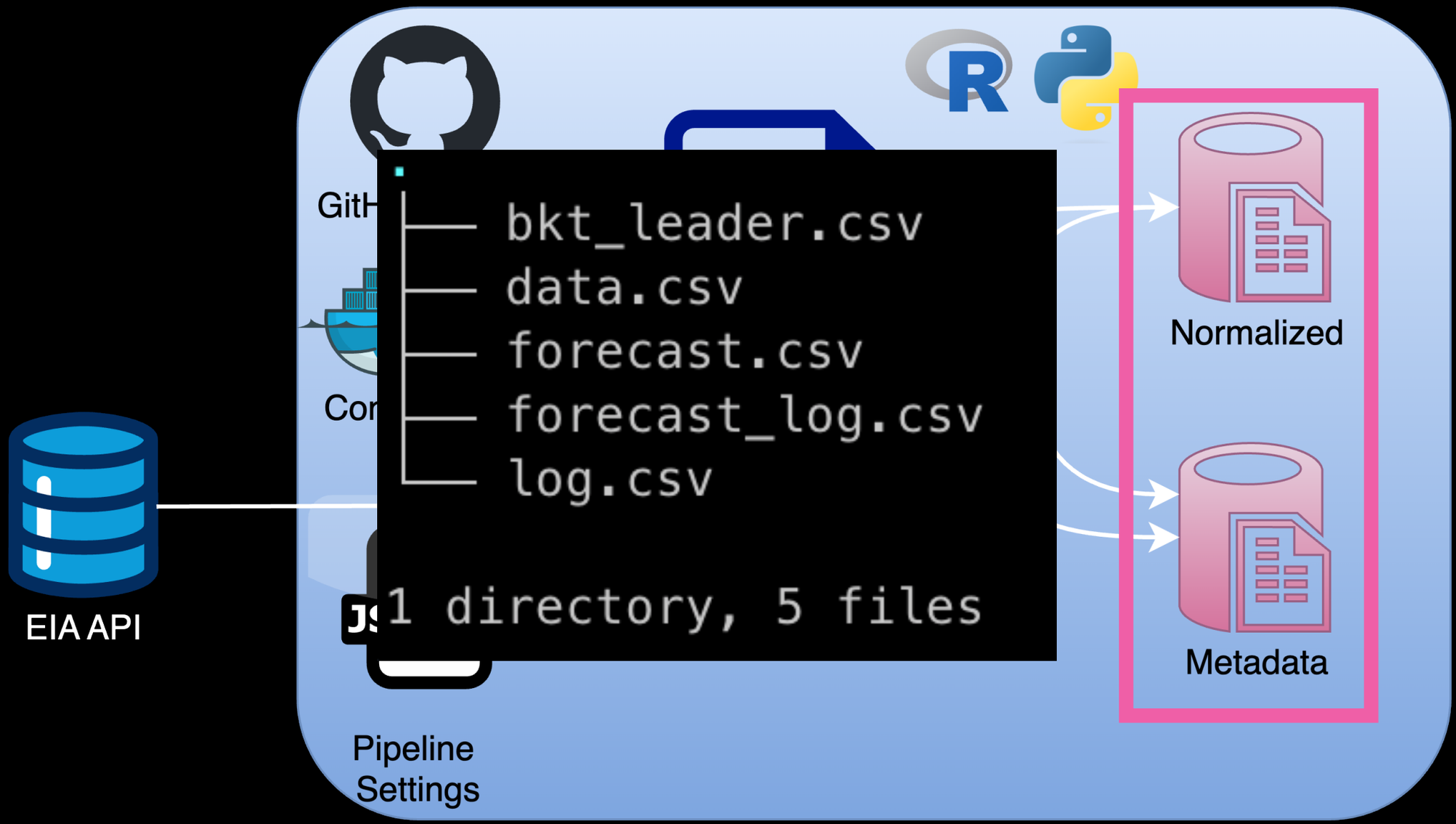
Preview | Code | Blame | 361 lines (361 loc) · 63.3 KB | Code 55% faster with GitHub Copilot | Raw

Search this file

respondent	respondent_type	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	type	update	success
US48	Demand	2024-06-04 09:09:29.847003	2015-07-01 05:00:00	2024-06-03 01:00:00	2015-07-01 05:00:00	2024-06-03 01:00:00	True	True	78237.0	0.0	backfill	True	True
US48	Demand	2024-06-04 09:13:35.486793	2024-06-03 02:00:00	2024-06-04 00:00:00	2024-05-27 02:00:00	2024-06-04 00:00:00	True	True	191.0	0.0	refresh	True	True
US48	Demand	2024-06-04 10:11:44.473285	2024-06-04 01:00:00	2024-06-04 01:00:00							refresh	False	False
US48	Demand	2024-06-04 11:10:51.547354	2024-06-04 01:00:00	2024-06-04 02:00:00	2024-05-28 01:00:00	2024-06-04 02:00:00	True	True	170.0	0.0	refresh	True	True
US48	Demand	2024-06-04 12:20:08.853848	2024-06-04 03:00:00	2024-06-04 03:00:00							refresh	False	False
US48	Demand	2024-06-04 13:13:27.580243	2024-06-04 03:00:00	2024-06-04 03:00:00							refresh	False	False
US48	Demand	2024-06-04 14:11:05.642034	2024-06-04 03:00:00	2024-06-04 23:00:00	2024-05-28 03:00:00	2024-06-04 11:00:00	True	False	177.0	0.0	refresh	True	True
US48	Demand	2024-06-04 15:12:08.108327	2024-06-04 12:00:00	2024-06-04 23:00:00	2024-05-28 12:00:00	2024-06-04 14:00:00	True	False	171.0	0.0	refresh	True	True
US48	Demand	2024-06-04 16:14:20.455581	2024-06-04 15:00:00	2024-06-04 23:00:00	2024-05-28 15:00:00	2024-06-04 15:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 17:10:30.755251	2024-06-04 16:00:00	2024-06-04 23:00:00	2024-05-28 16:00:00	2024-06-04 16:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 18:14:45.830771	2024-06-04 17:00:00	2024-06-04 23:00:00	2024-05-28 17:00:00	2024-06-04 17:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 19:09:33.113685	2024-06-04 18:00:00	2024-06-04 23:00:00	2024-05-28 18:00:00	2024-06-04 18:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 20:12:00.866350	2024-06-04 19:00:00	2024-06-04 23:00:00	2024-05-28 19:00:00	2024-06-04 19:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 21:10:53.584028	2024-06-04 20:00:00	2024-06-04 23:00:00	2024-05-28 20:00:00	2024-06-04 20:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 22:11:36.357769	2024-06-04 21:00:00	2024-06-04 23:00:00	2024-05-28 21:00:00	2024-06-04 21:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-04 23:12:31.360776	2024-06-04 22:00:00	2024-06-04 23:00:00	2024-05-28 22:00:00	2024-06-04 22:00:00	True	False	169.0	0.0	refresh	True	True
US48	Demand	2024-06-05 00:48:57.513133	2024-06-04 23:00:00	2024-06-04 23:00:00							refresh	False	False

Pipeline Design

Not the best practice!



Pipeline Design

EIA API - Data Refrsh (Python Version)

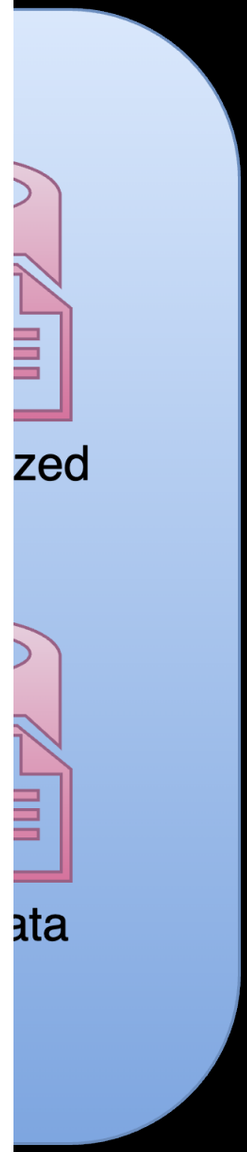
Load libraries

```
import eia_api as api
import eia_data
import pandas as pd
import numpy as np
import requests
import json
import os
import datetime
import plotly.express as px
import great_tables as gt
```

API Settings:

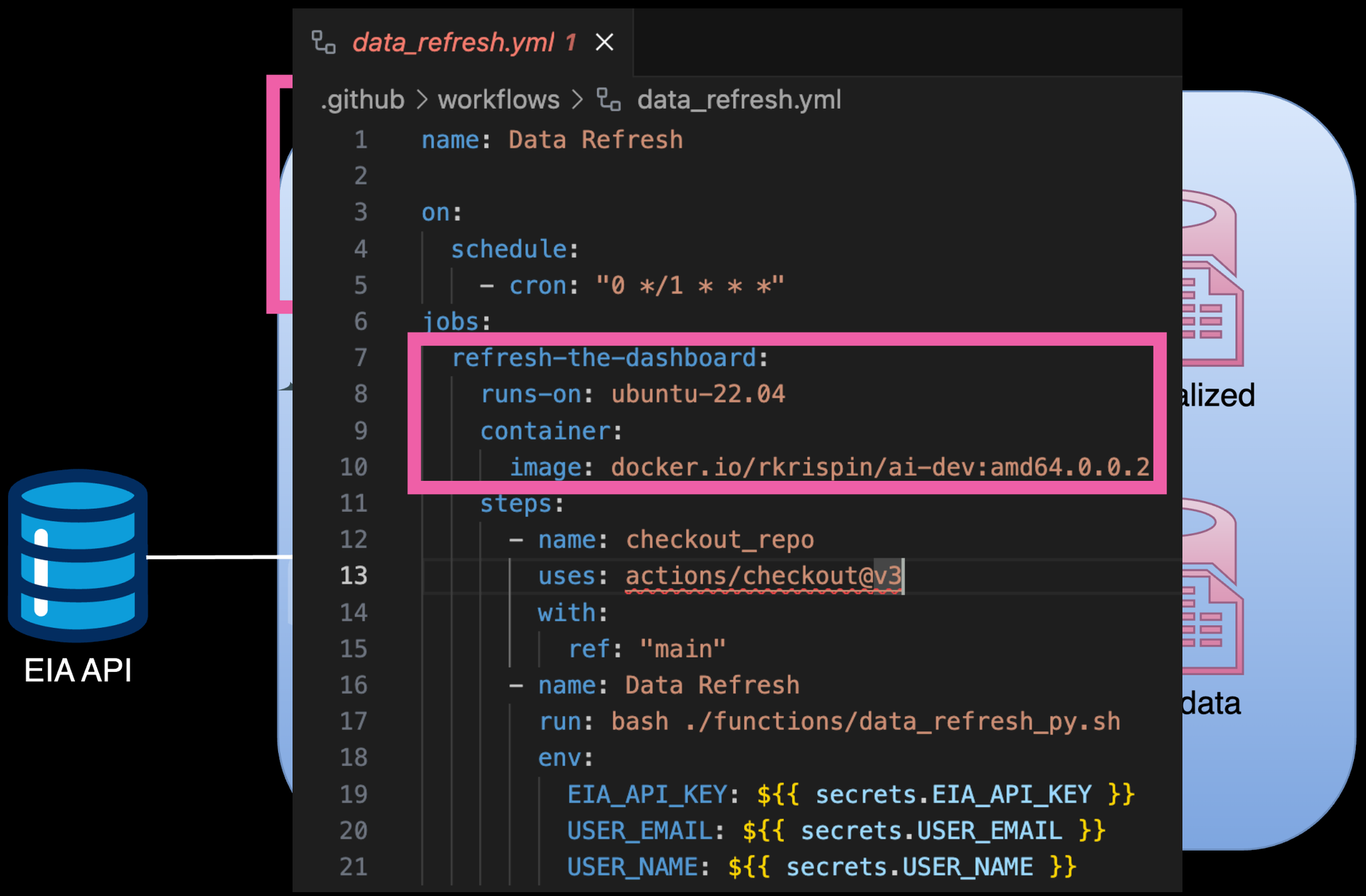
```
raw_json = open("../settings/series.json")
meta_json = json.load(raw_json)
series = pd.DataFrame(meta_json["series"])
api_path = meta_json["api_path"]

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

offset = 2250

eia_api_key = os.getenv('EIA_API_KEY')

meta_path = meta_json["meta_path"]
data_path = meta_json["data_path"]
```

```
meta_obj = eia_data.get_metadata(api_key = eia_api_key, api_path = api_path, meta_path =

gt.GT(meta_obj.request_meta,)
```

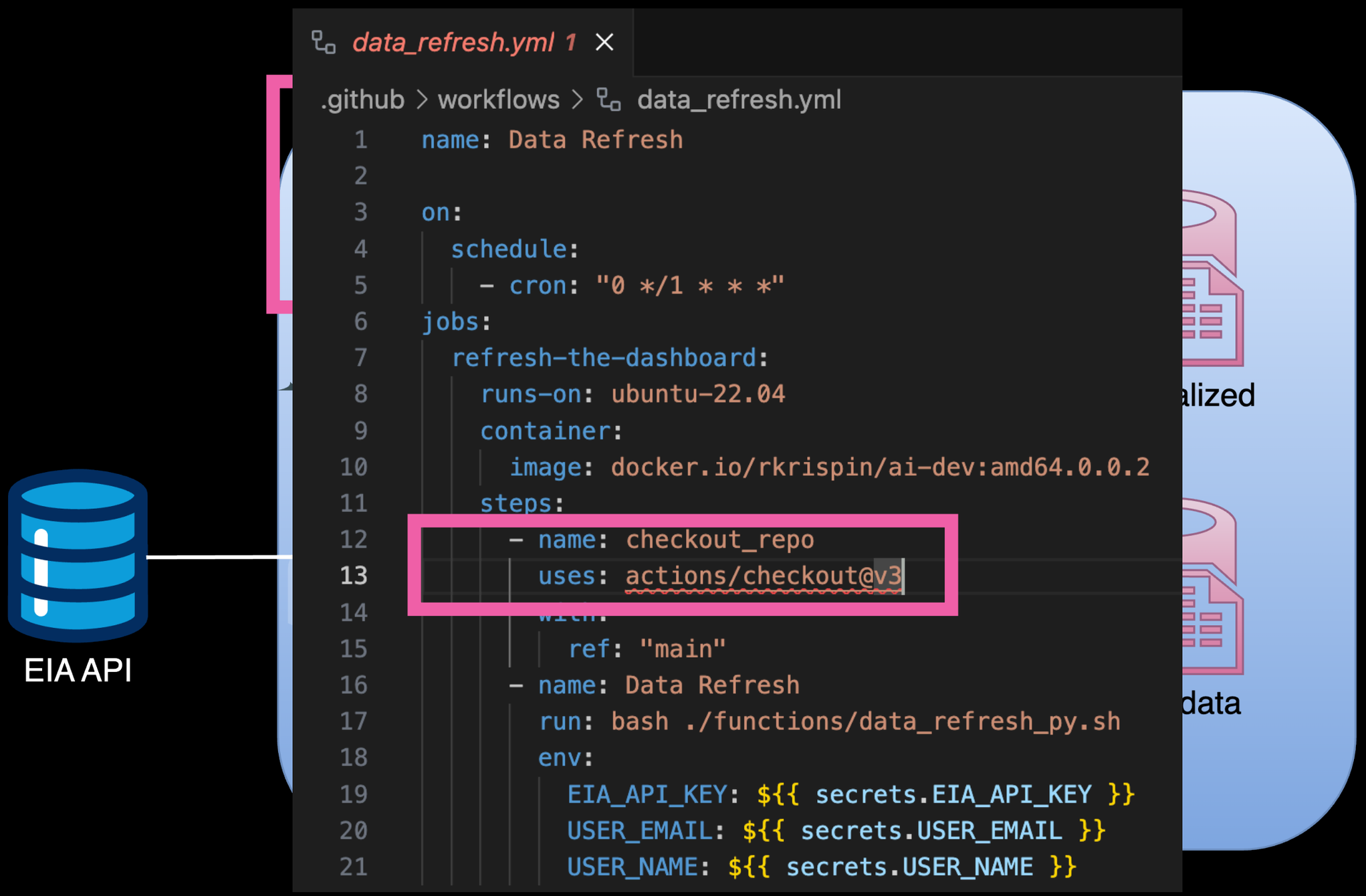


Pipeline Design

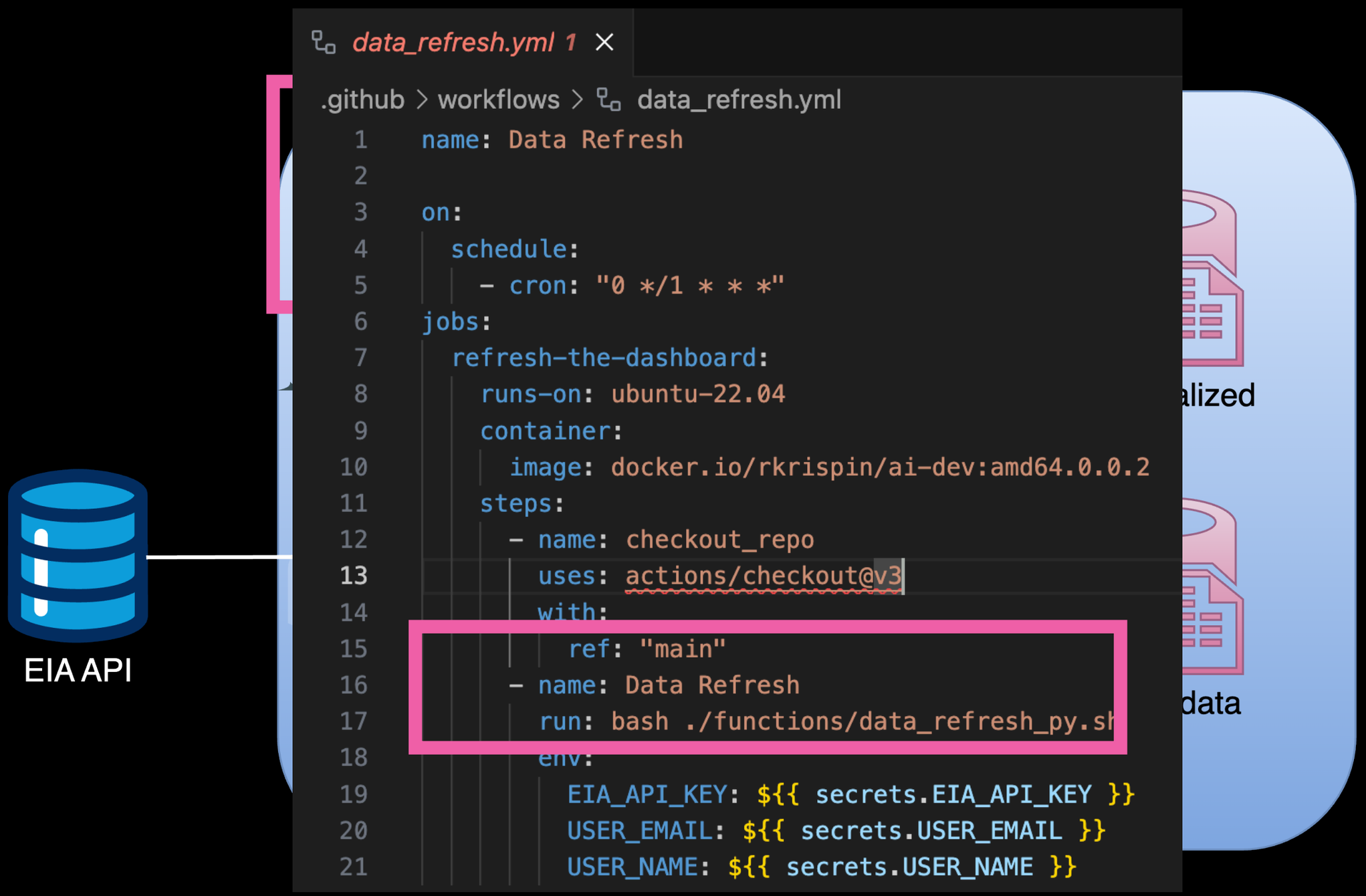
Pipeline Design



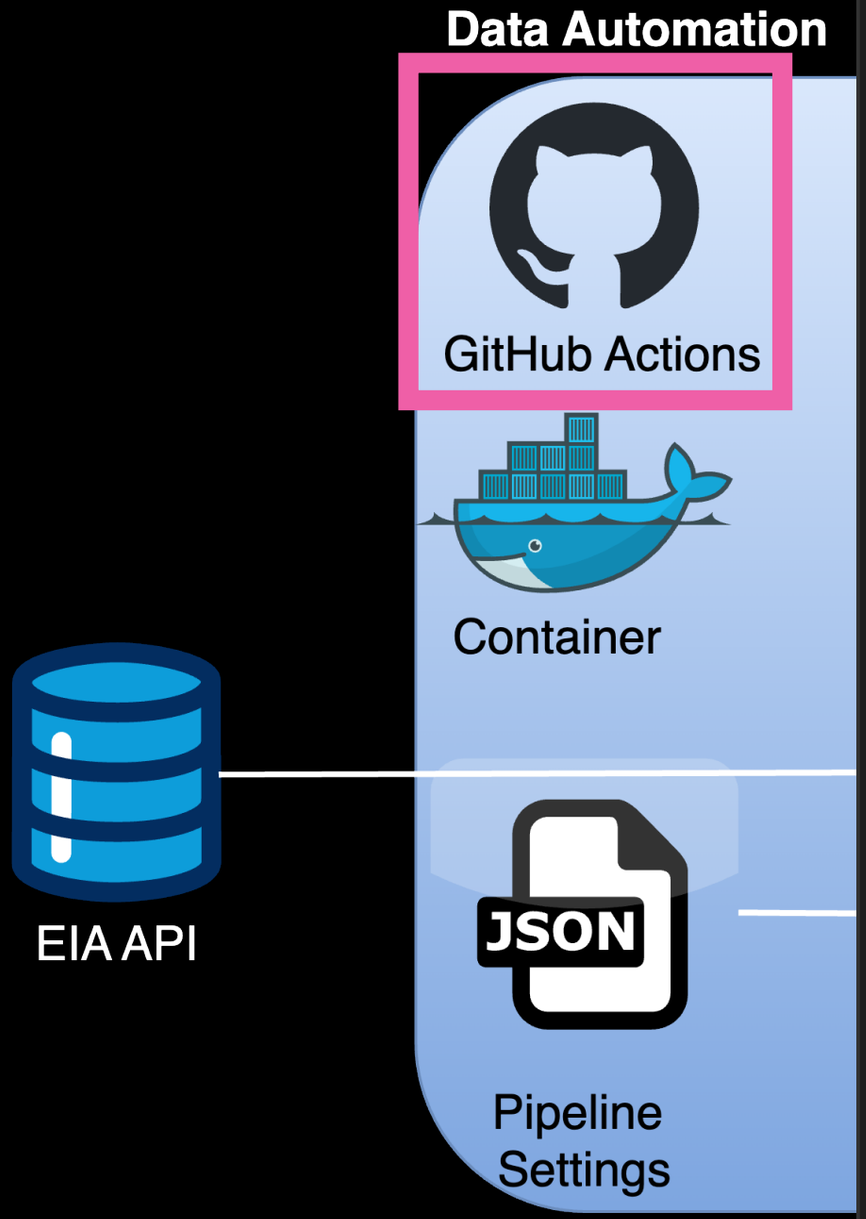
Pipeline Design



Pipeline Design



Pipeline Desi

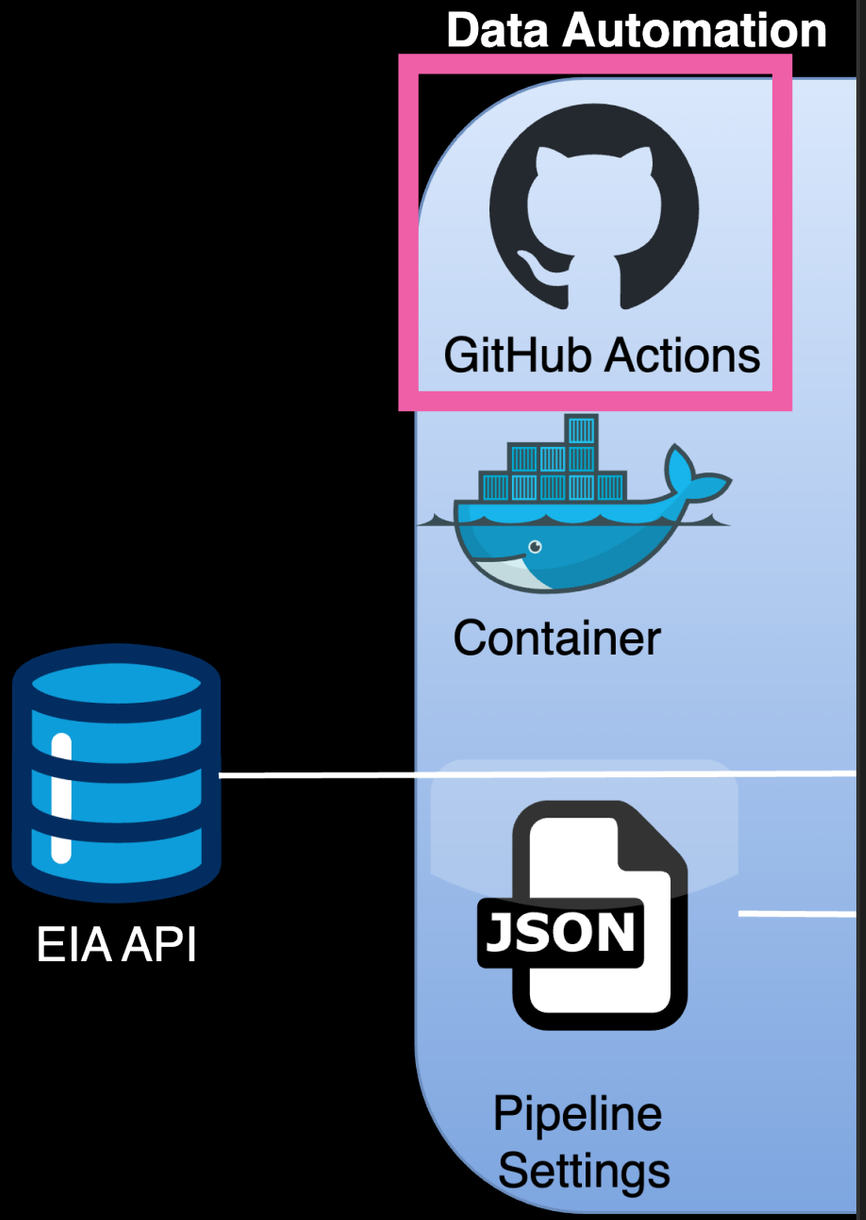


$ data_refresh_py.sh

functions > $ data_refresh_py.sh

```
#!/usr/bin/env bash
source /opt/$VENV_NAME/bin/activate

rm -rf ./functions/data_refresh_py_files
rm -rf ./functions/data_refresh_py.html
quarto render ./functions/data_refresh_py.qmd --to html

rm -rf docs/data_refresh/
mkdir docs/data_refresh
cp ./functions/data_refresh_py.html ./docs/data_refresh/
cp -R ./functions/data_refresh_py_files ./docs/data_refresh/

echo "Finish"
p=$(pwd)
git config --global --add safe.directory $p


# Render the Quarto dashboard
if [[ "$(git status --porcelain)" != "" ]]; then
    quarto render functions/index.qmd
    cp functions/index.html docs/index.html
    rm -rf docs/index_files
    cp -R functions/index_files/ docs/
    rm functions/index.html
    rm -rf functions/index_files
    git config --global user.name $USER_NAME
    git config --global user.email $USER_EMAIL
    git add data/*
    git add docs/*
    git commit -m "Auto update of the data"
    git push origin main
else
echo "Nothing to commit..."
fi
```

Pipeline Desi

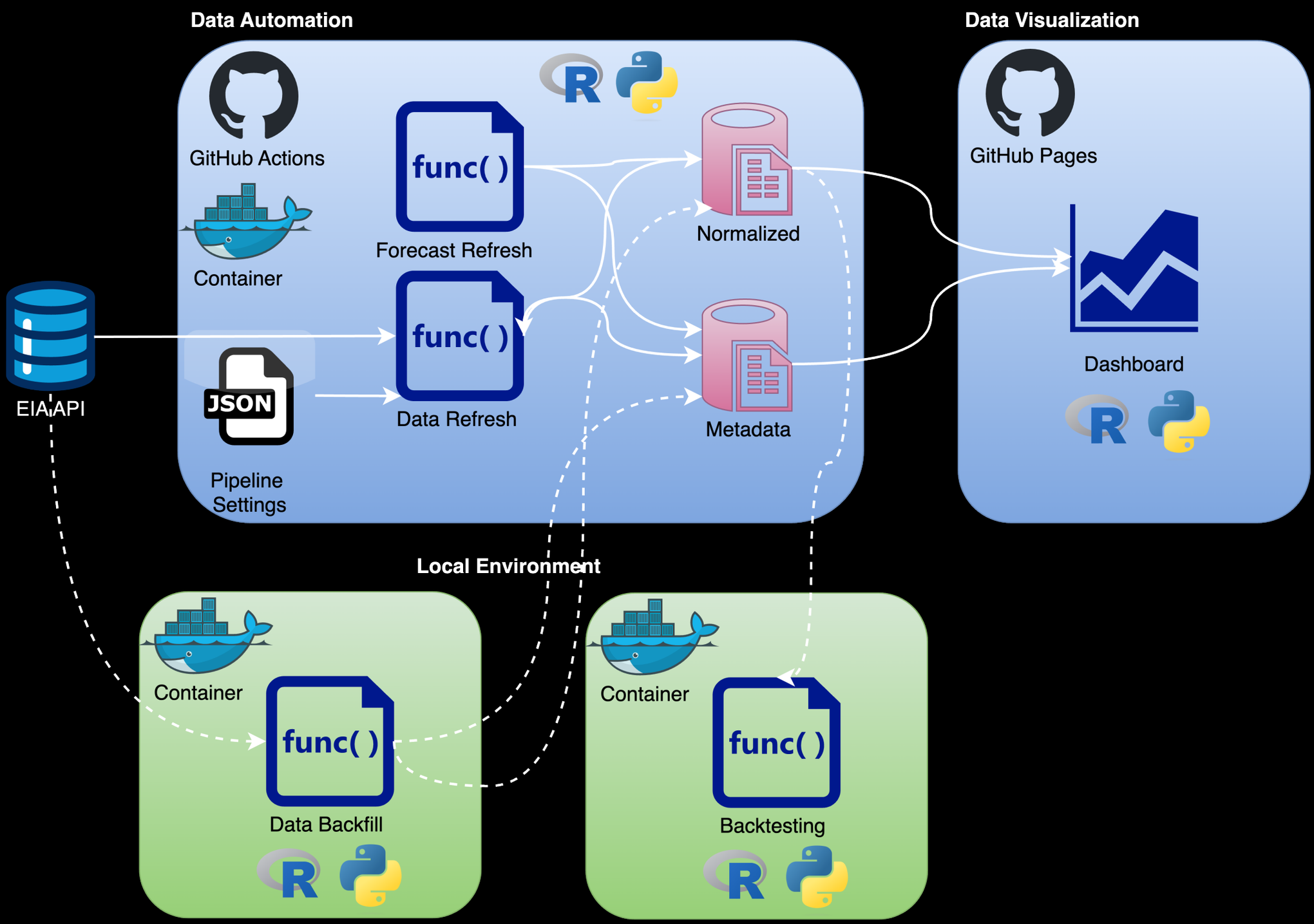


$ data_refresh_py.sh

functions > $ data_refresh_py.sh

```
#!/usr/bin/env bash
source /opt/$VENV_NAME/bin/activate

rm -rf ./functions/data_refresh_py_files
rm ./functions/data_refresh_py.html
quarto render ./functions/data_refresh_py.qmd --to html

rm -rf docs/data_refresh/
mkdir docs/data_refresh
cp ./functions/data_refresh_py.html ./docs/data_refresh/
cp -R ./functions/data_refresh_py_files ./docs/data_refresh/

echo "Finish"
p=$(pwd)
git config --global --add safe.directory $p

# Render the Quarto dashboard
if [[ "$(git status --porcelain)" != "" ]]; then
    quarto render functions/index.qmd
    cp functions/index.html docs/index.html
    rm -rf docs/index_files
    cp -R functions/index_files/ docs/
    rm functions/index.html
    rm -rf functions/index_files
    git config --global user.name $USER_NAME
    git config --global user.email $USER_EMAIL
    git add data/*
    git add docs/*
    git commit -m "Auto update of the data"
    git push origin main
else
echo "Nothing to commit..."
fi
```

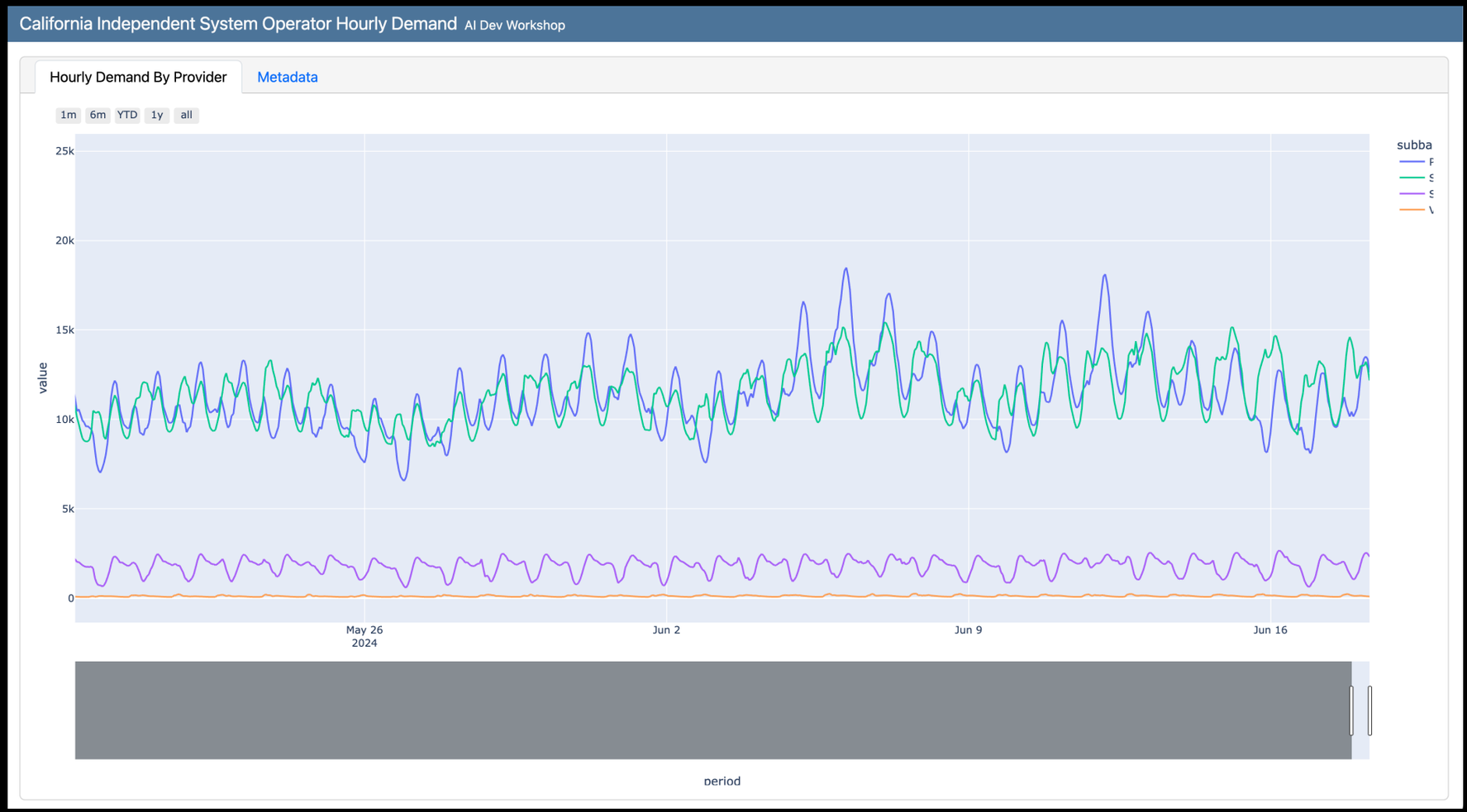
Pipeline Design



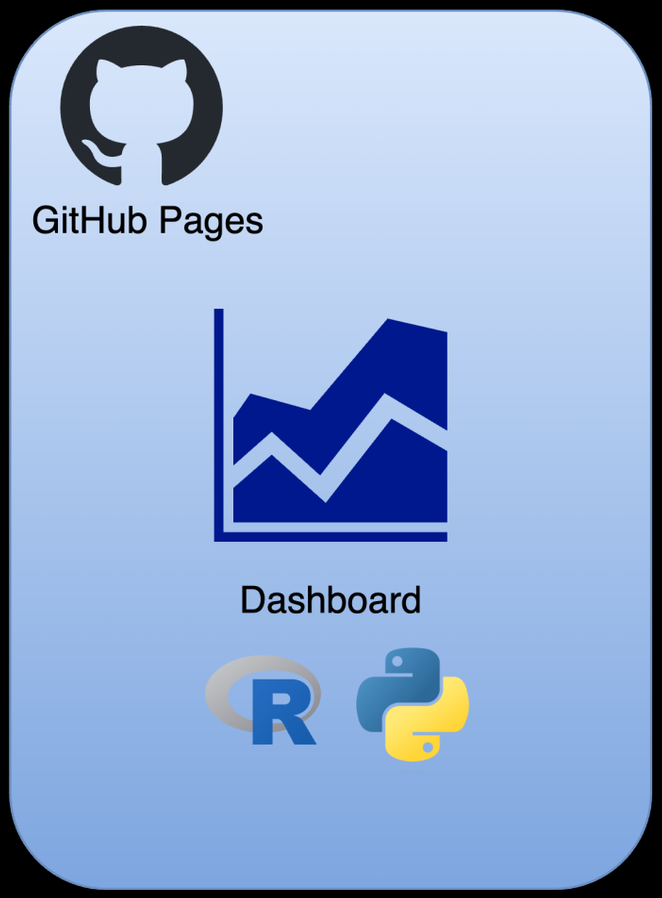
Pipeline Design

GitHub Pages

Dashboard



Pipeline Design



California Independent System Operator Hourly Demand AI Dev Workshop

Hourly Demand By Provider | Metadata

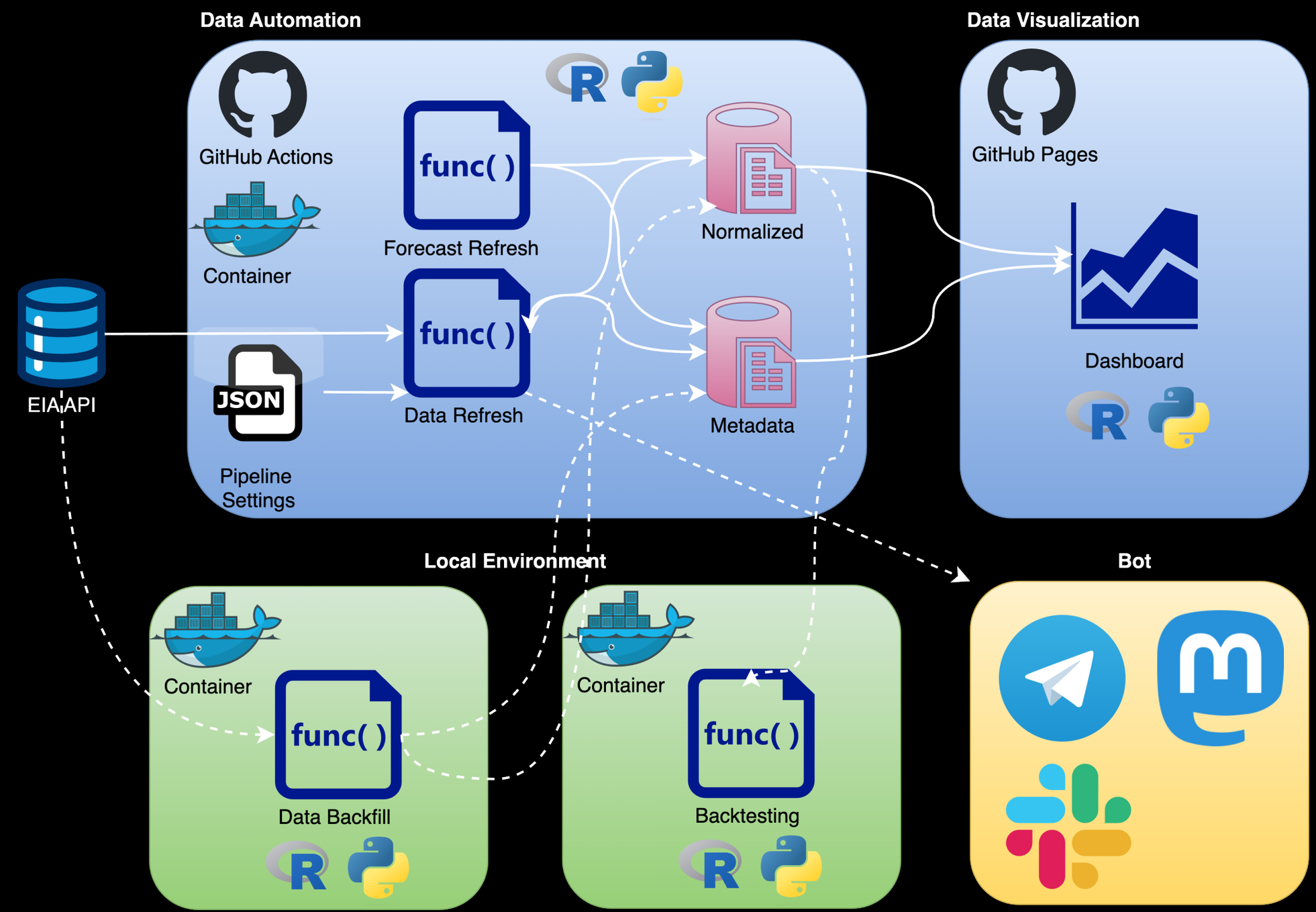
index	parent	subba	time	start	end	start_act	end_act	start_match	end_match	n_obs	na	type	update	success	comments
41	CISO	PGAE	2024-06-19 04:19:47.237844+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
41	CISO	SCE	2024-06-19 04:19:47.238733+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
41	CISO	SDGE	2024-06-19 04:19:47.240187+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
41	CISO	VEA	2024-06-19 04:19:47.241557+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
42	CISO	PGAE	2024-06-19 05:16:48.928662+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
42	CISO	SCE	2024-06-19 05:16:48.929611+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
42	CISO	SDGE	2024-06-19 05:16:48.931054+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
42	CISO	VEA	2024-06-19 05:16:48.932411+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
43	CISO	PGAE	2024-06-19 06:22:25.443620+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
43	CISO	SCE	2024-06-19 06:22:25.444909+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
43	CISO	SDGE	2024-06-19 06:22:25.446762+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
43	CISO	VEA	2024-06-19 06:22:25.448279+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
44	CISO	PGAE	2024-06-19 07:15:18.244598+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;
44	CISO	SCE	2024-06-19 07:15:18.245507+00:00	2024-06-18 08:00:00	2024-06-18 07:00:00							refresh	False	False	No new data is available; The data refresh failed, please check the log;

Demo

What else can we do?

Many other cool things!

Adding Bot


Data Automation
Data Visualization
GitHub Actions
Container
Forecast Refresh
Normalized
EIA API
JSON
Data Refresh
Metadata
Pipeline Settings
GitHub Pages
Dashboard
Local Environment
Bot
Container
Data Backfill
Container
Backtesting

Best Practices

- 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