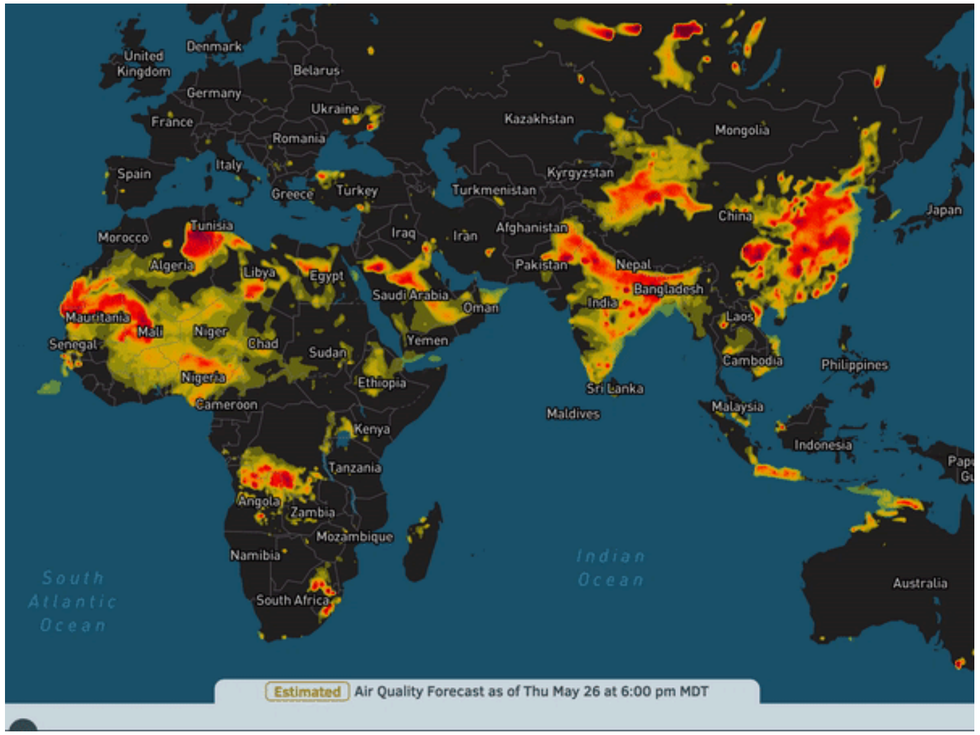
BAQ

BANGKOK AIR QUALITY

INTRODUCTION

Air pollution, particularly **PM2.5**, has significant impacts on public health and the environment. Predicting PM2.5 levels may help authorities and individuals take proactive measures to reduce exposure.

INTRODUCTION



This project aims to develop a machine learning model to forecast PM2.5 concentrations using historical air quality and weather data in Bangkok.
By providing timely and accurate forecasts, the system can support public health interventions and help people make informed decisions to reduce health risks.

SCOPE & CONSTRAINT

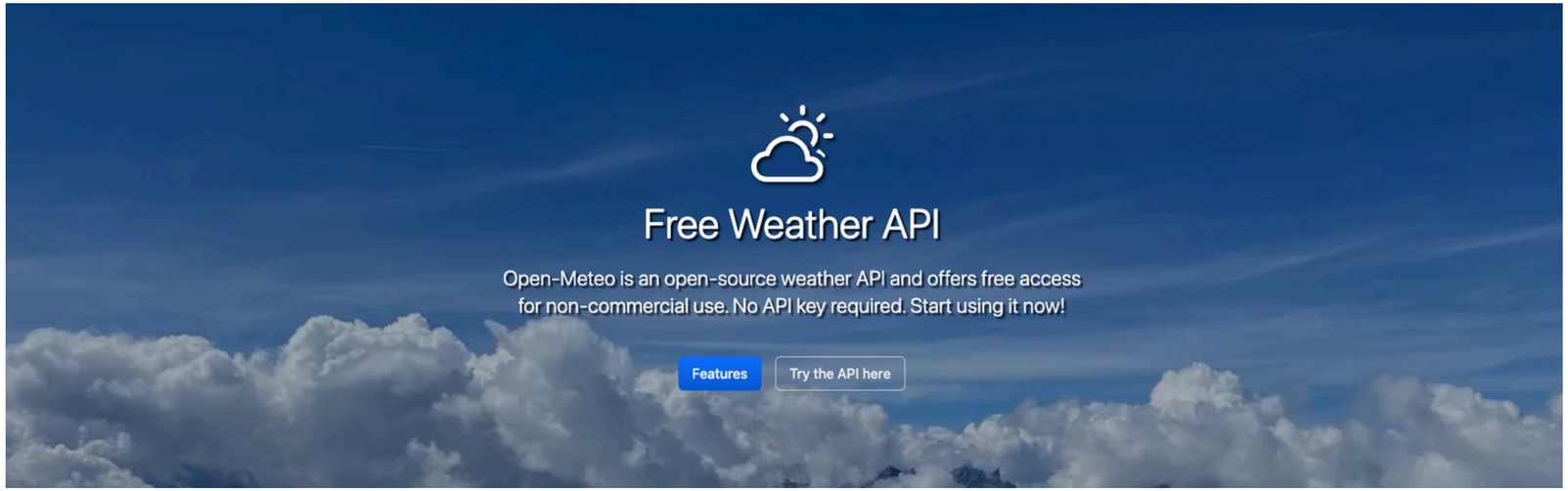
This project focuses on building a complete MLOps pipeline for forecasting PM2.5 concentrations in Bangkok.
The scope includes

- Data Pipeline
  - Develop an automated data pipeline to collect real-time and historical air quality and weather data (e.g., PM2.5, PM10, humidity, temperature) from the Open-Meteo API
- Model Development
  - Train time-series forecasting models using machine learning (Random Forest, XGBoost), and deep learning (LSTM) to predict air pollution levels.
- Deployment and Monitoring
  - Deploy the model via FastAPI for real-time predictions and build an interactive web application to display forecasts. Implement CI/CD pipelines for automated testing, versioning, and deployment.

SCOPE & CONSTRAINT

This project focuses on building a complete MLOps pipeline for forecasting PM2.5 concentrations in Bangkok.
The Constraints are

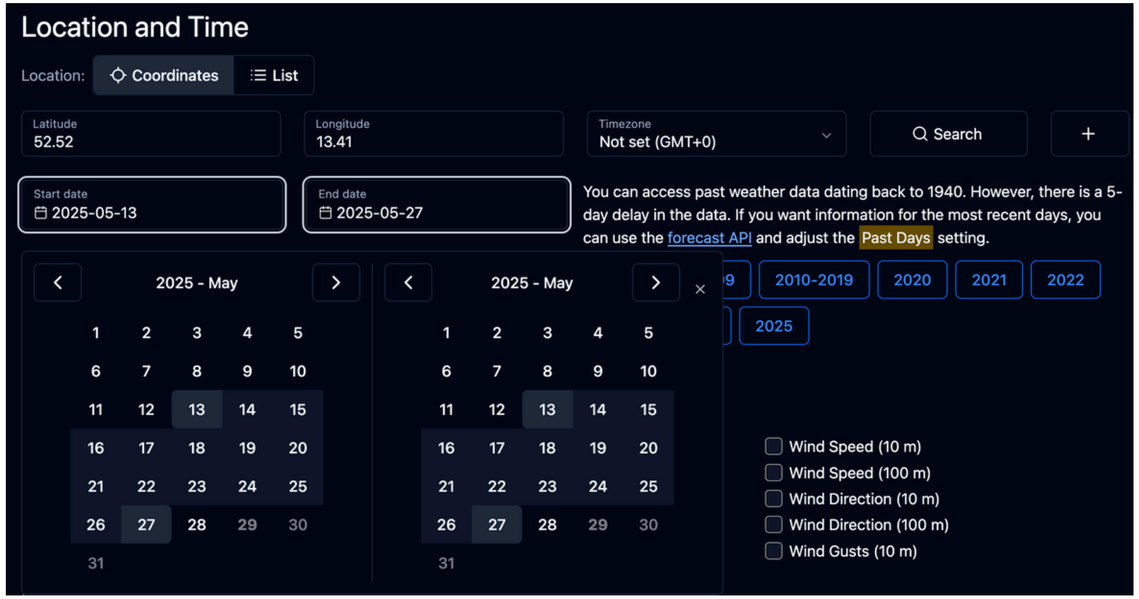
- Data Availability and Quality :
  - The historical weather data from OpenMeteo used for model training is itself forecasted data with periodic validation, rather than actual observed data.
  This may affect the forecasting model's performance.



SCOPE & CONSTRAINT

This project focuses on building a complete MLOps pipeline for forecasting PM2.5 concentrations in Bangkok.
The Constraints are

- Delay in Access to Recent Data
  - The most recent data available from Open-Meteo typically have a delay by 24–72 hours.
    This delay may limits the model's ability to incorporate the most current environmental conditions, which may reduce the result of short-term forecasting or real-time responsiveness.



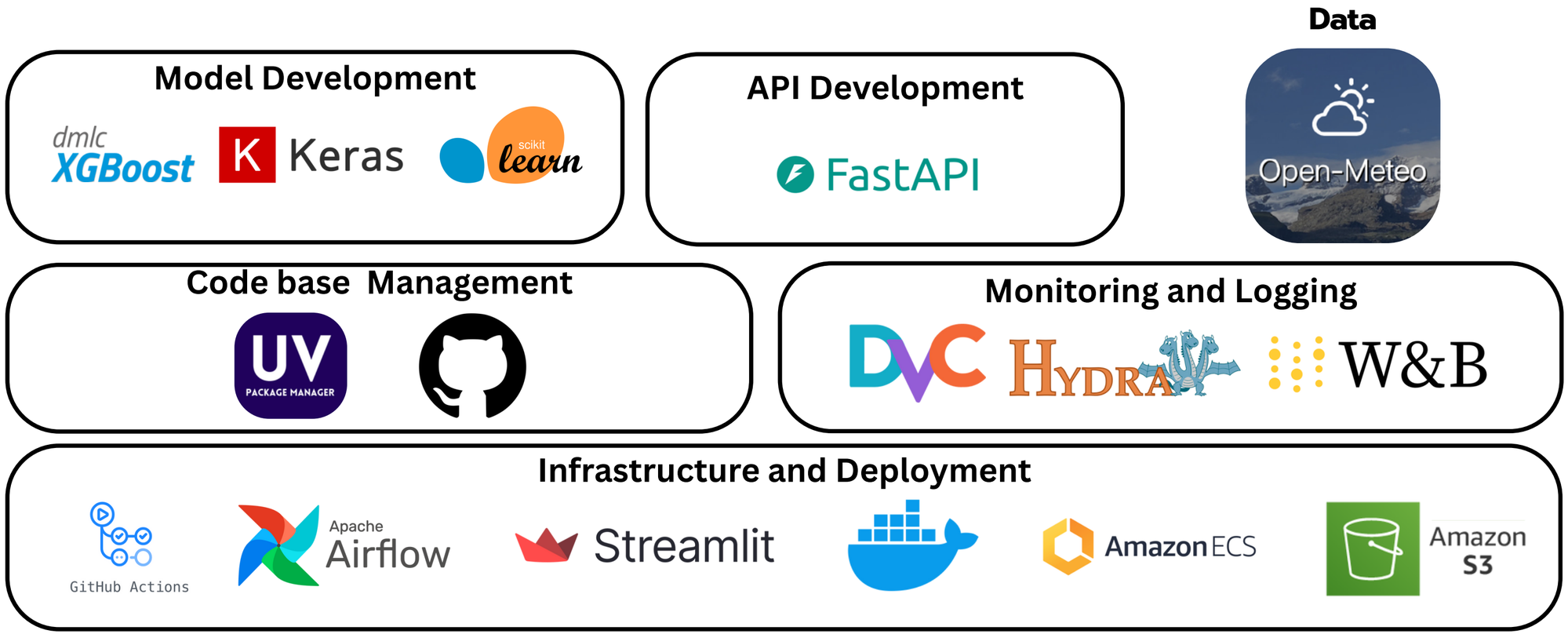
This project focuses on building a complete MLOps pipeline for forecasting PM2.5 concentrations in Bangkok.
The Constraints are

- Budget Constraint
  - Due to financial constraints, the team aims to minimize costs by prioritizing the use of free or open-source tools and platforms wherever possible, without compromising project quality.
- Use of AWS Credits
  - The project infrastructure is deployed on AWS, as the team already has existing AWS credits. This enables the project to access scalable cloud resources while avoiding additional infrastructure costs.

STAKEHOLDER

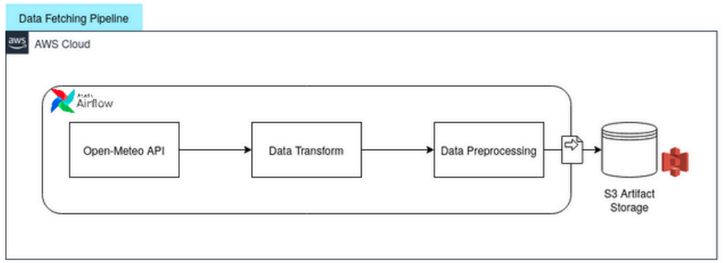
Stakeholders	Needs
Bangkok Residents	An accurate and timely PM2.5 forecast to guide their daily activities
Academic Advisors	Supervise the scientific rigor of the project, make sure the technique complies with academic standards, and offer advice on data analysis and model validation.
Project Team (Data Scientists & MLOps Engineers)	Responsible for the entire lifecycle. From data collection, preprocessing, and model development to deployment and monitoring and maintenance of the forecasting system.

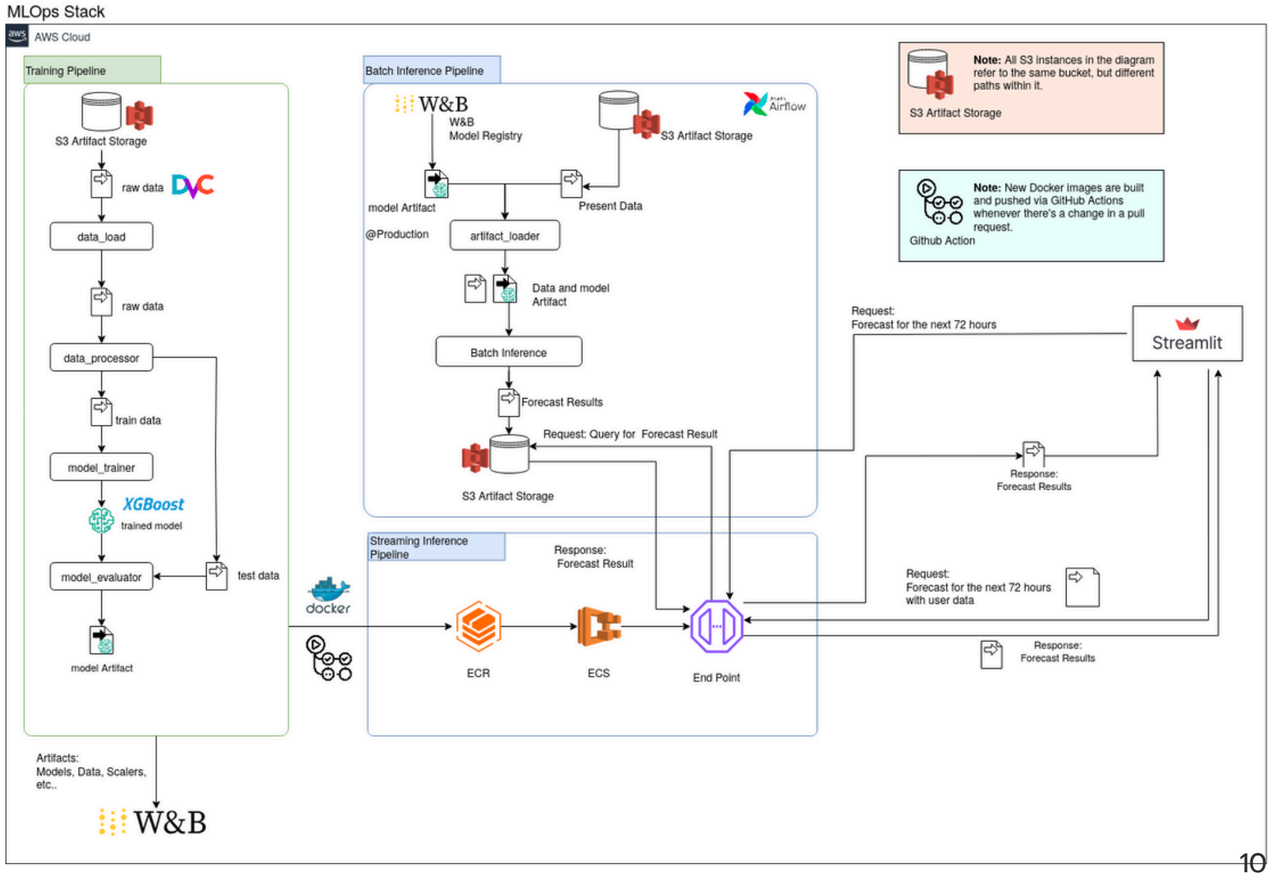
ML & Data Management



Architecture Design

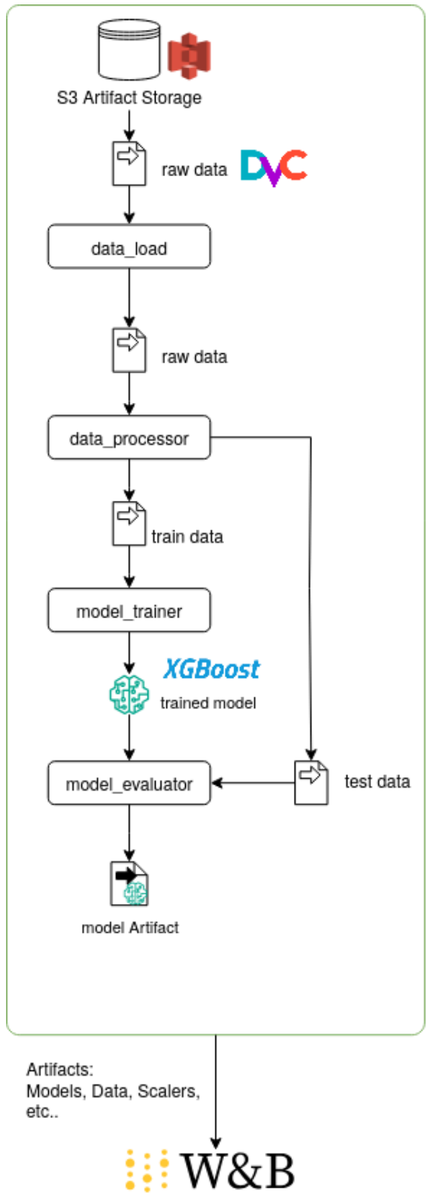
This MLOps architecture covers data fetching, training, batch inference, and streaming inference, leveraging Airflow, S3, Docker, and Streamlit for a comprehensive machine learning pipeline.





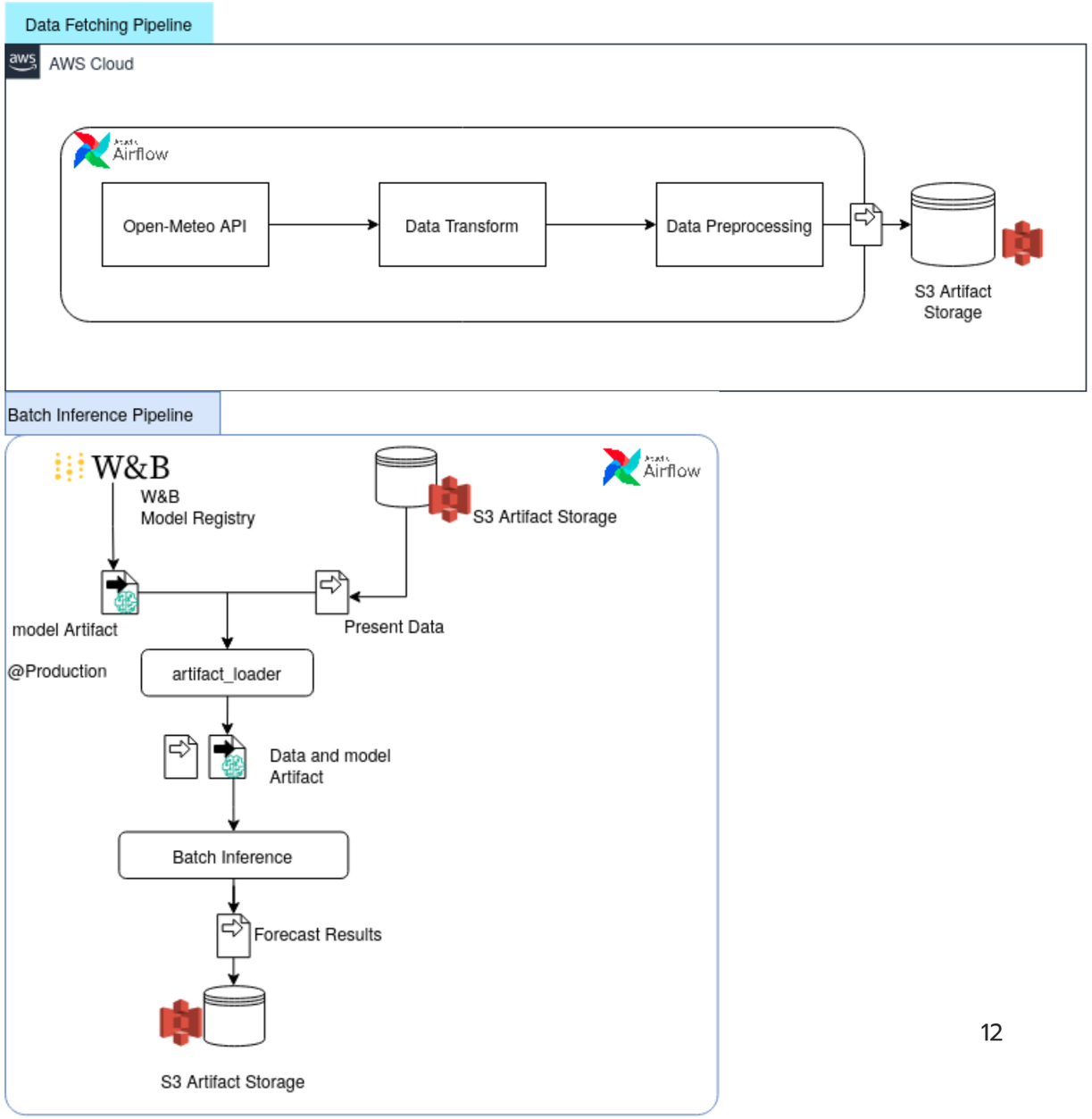
Training Pipeline

This orchestrator-agnostic ML pipeline loads raw data from S3 (tracked by DVC), processes it, trains an XGBoost model, evaluates it, and tracks all artifacts using Weights & Biases.



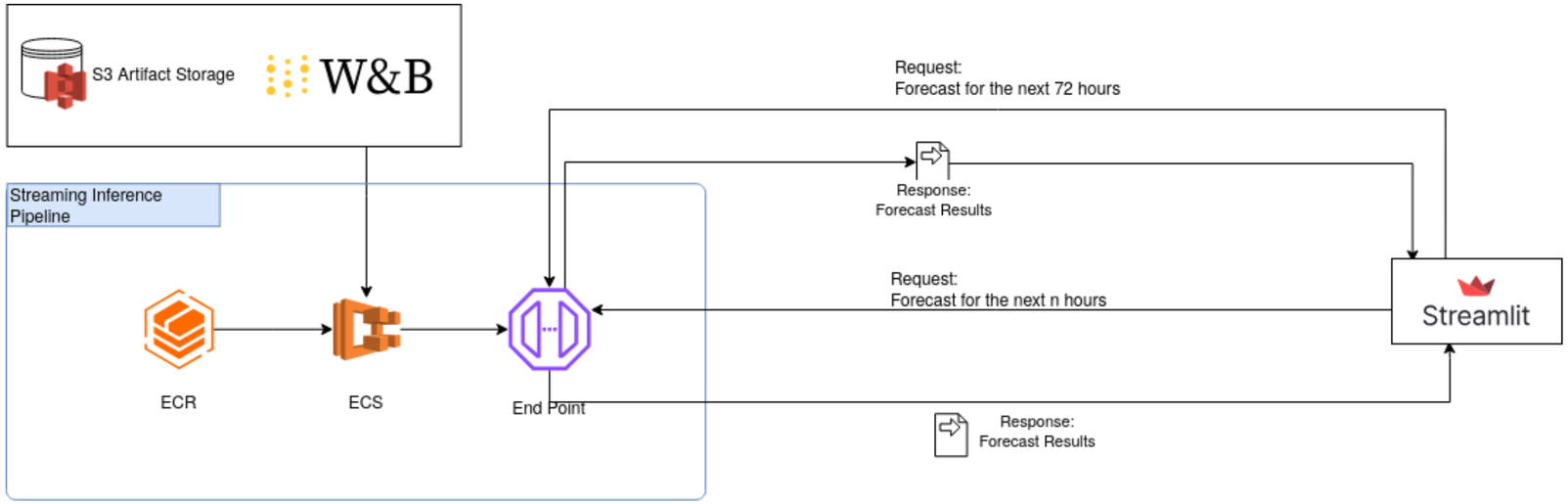
Airflow

1. **Data Fetching Pipeline:** Fetches data from Open-Meteo API, transforms, preprocesses it, and stores the processed data in S3 Artifact Storage, all orchestrated by Airflow.
2. **Batch Inference Pipeline**: Retrieves model artifacts from W&B Model Registry and data from S3, loads them, performs batch inference, and saves the forecast results to S3 Artifact Storage, also managed by Airflow.

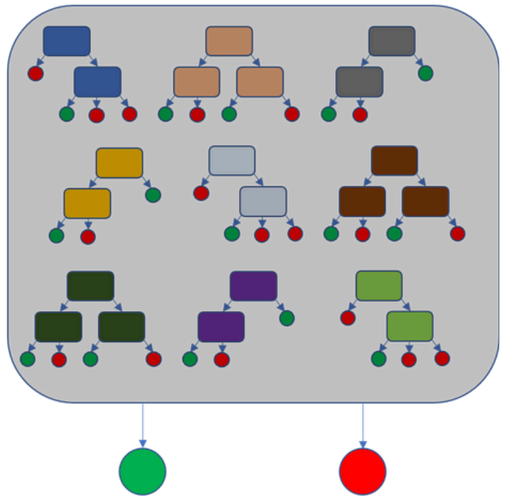


API

The endpoint supports both real-time and batch inference. It fetches live forecasts via ECS-based streaming services or retrieves precomputed results from S3, generated by batch inference DAGs. Streamlit queries the endpoint to display forecasts for user-defined time ranges.

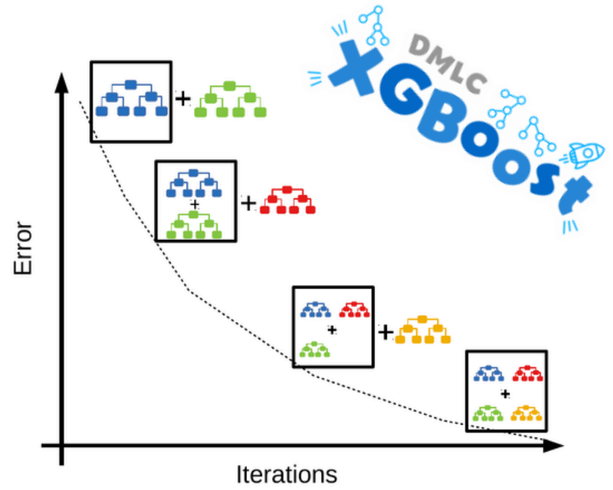


MODEL SELECTION



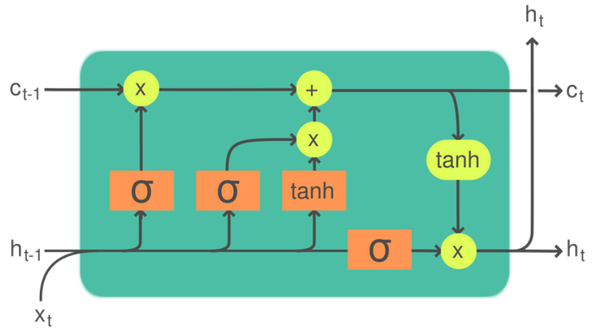
RANDOM FOREST

An ensemble of decision trees that improves accuracy and reduces overfitting.



XGBOOST

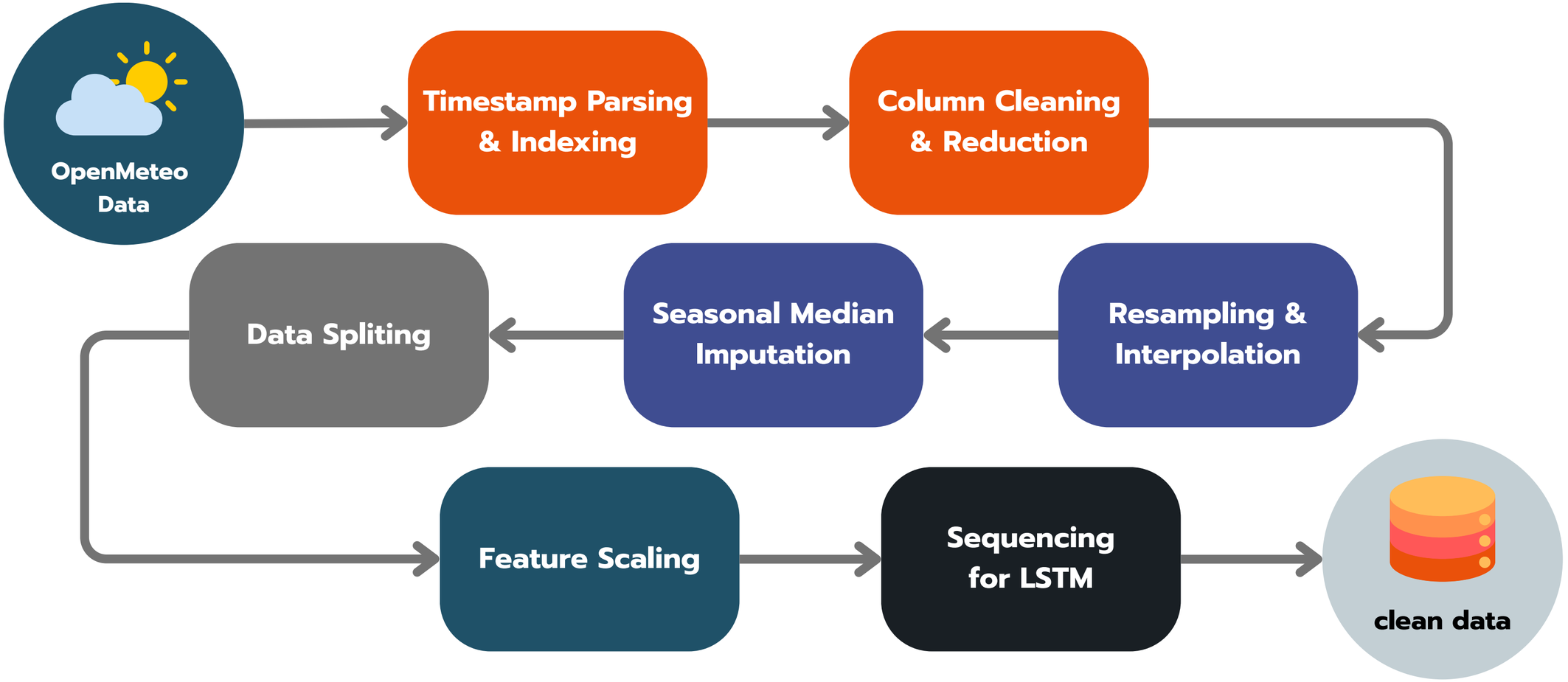
A fast, high-performance boosting algorithm that builds trees sequentially.



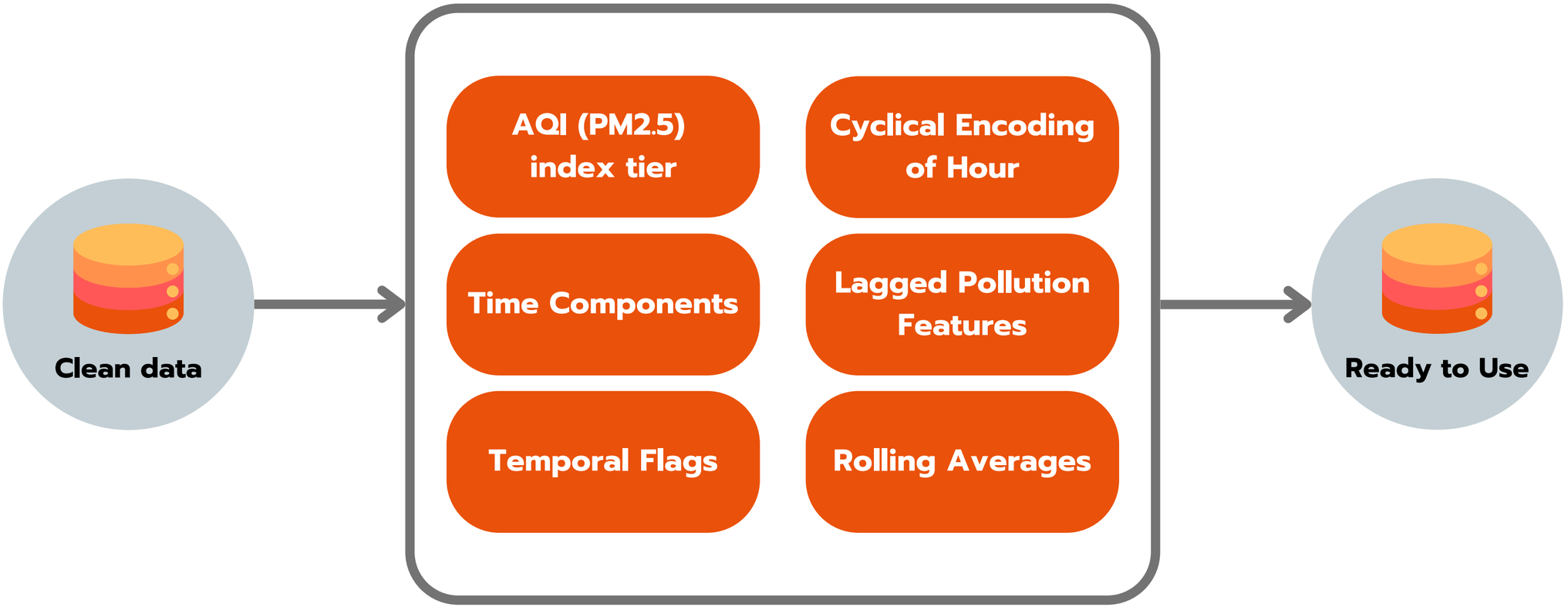
LSTM

A neural network for time series that learns from sequential patterns and long-term dependencies.

MODEL DEVELOPMENT - DATA PREPROCESSING



MODEL DEVELOPMENT - FEATURE ENGINEERING

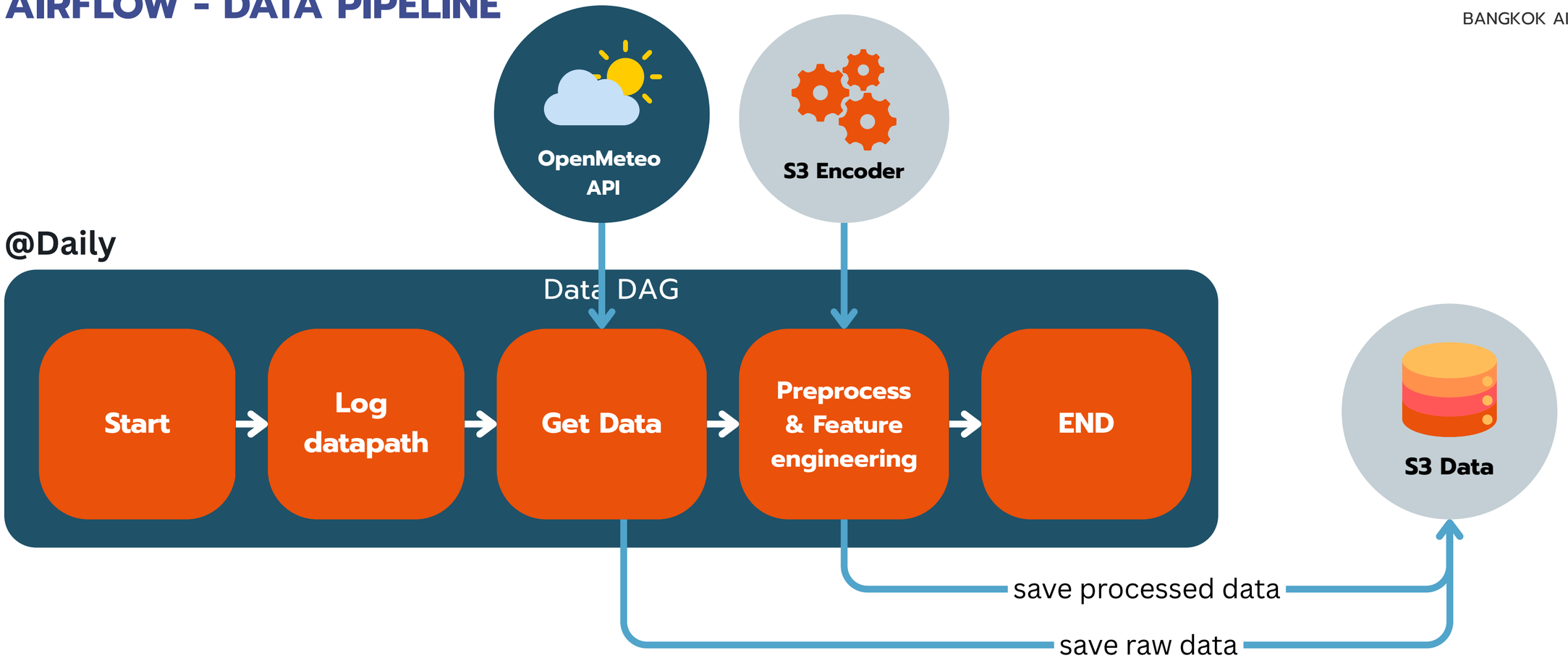


MODEL DEVELOPMENT

EVALUATION RESULT

MODEL	FORECAST TYPE	ACCURACY	MSE	RMSE	MAE	MAPE	R-SQUARED
XGBoost	Single-Step	0.9345	0.00027	0.01648	0.00745	0.0655	0.9640
	Multi-Step	0.9167	0.00006	0.00744	0.00609	0.0833	0.862
Random Forest	Single-Step	0.9309	0.00030	0.01740	0.00813	0.0691	0.9598
	Multi-Step	0.9715	0.00001	0.00283	0.00218	0.0285	0.9800
LSTM	Single-Step	0.8000	0.00077	0.02780	0.02040	0.2000	0.8983
	Multi-Step	0.9110	0.00006	0.00739	0.00666	0.0890	0.8639

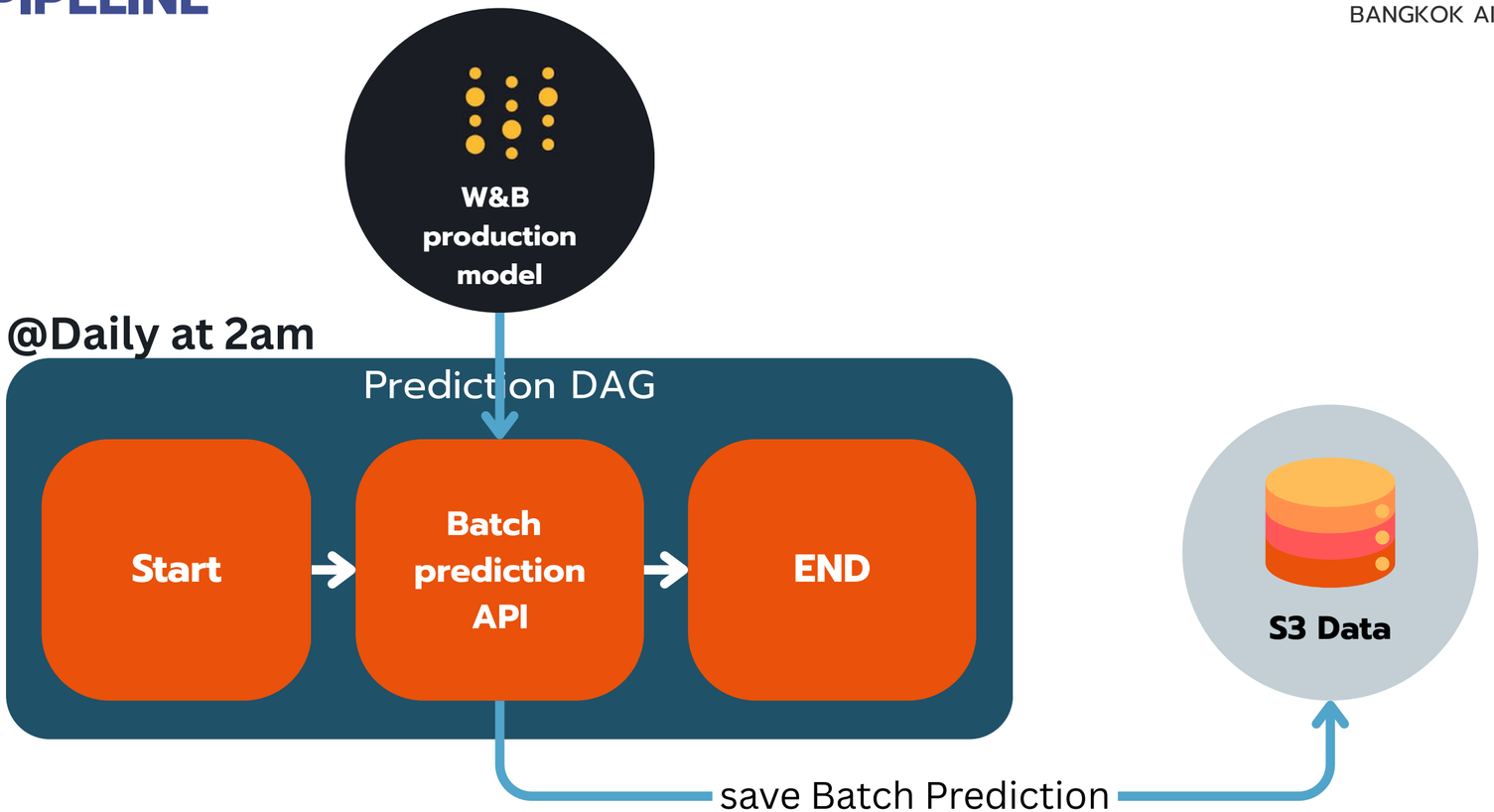
AIRFLOW - DATA PIPELINE



Note:

- **Custom Encoder and Scaler, stored in S3, are utilized in preprocessing process**

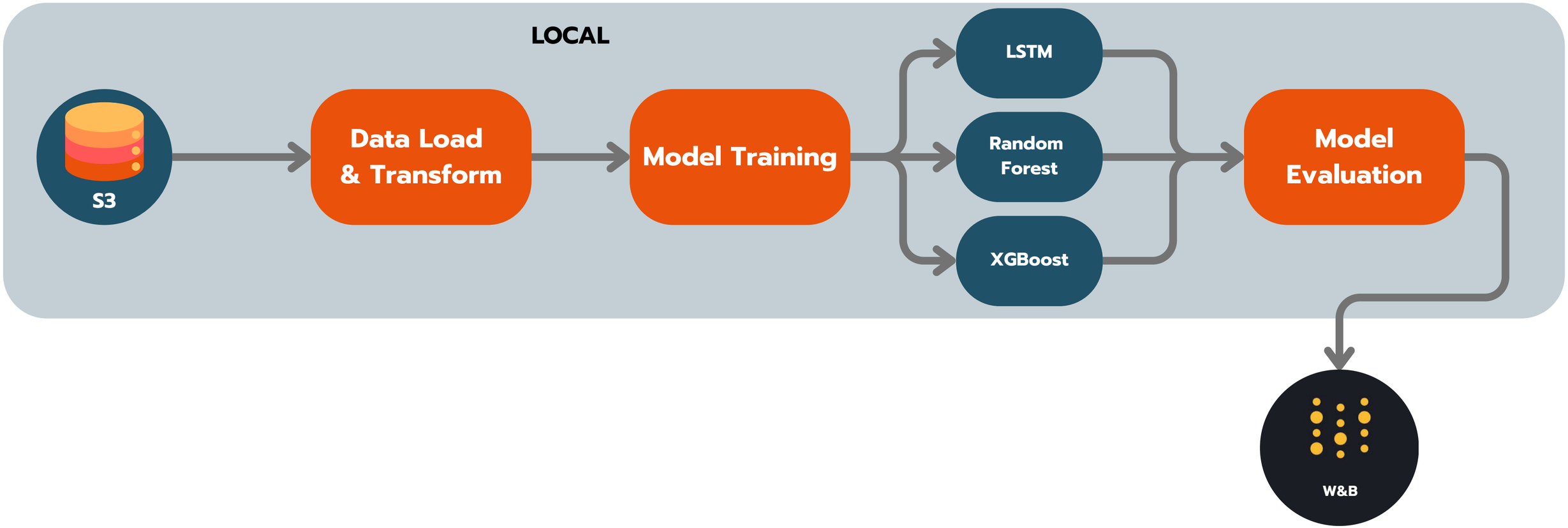
AIRFLOW - PREDICTION PIPELINE



Note:

- **The production stage Model in W&B will be used in prediction.**
- **Batch prediction will use one of the FAST-API endpoint.**

ML TRAINING PIPELINE



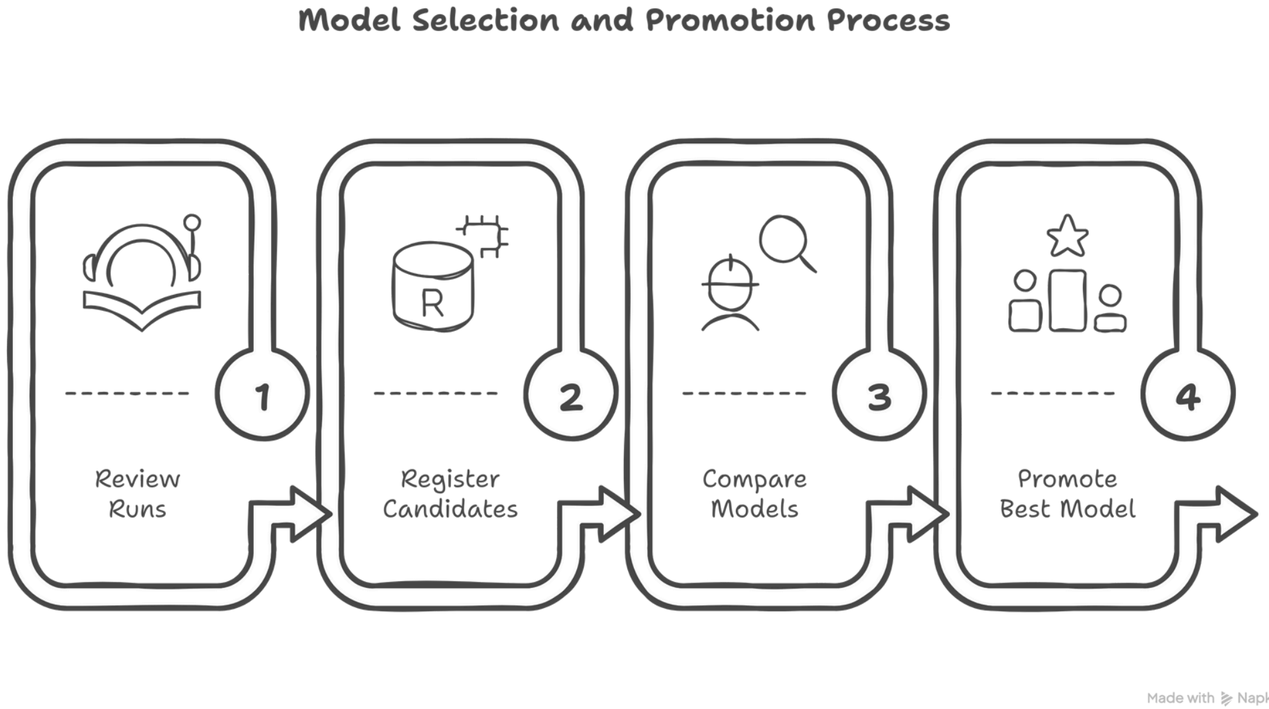
CI/CD with Github Action

We use GitHub Actions for CI/CD, automating model registration and promotion. This streamlines our workflow, ensuring efficient and reliable deployment of new models.

GitHub Actions

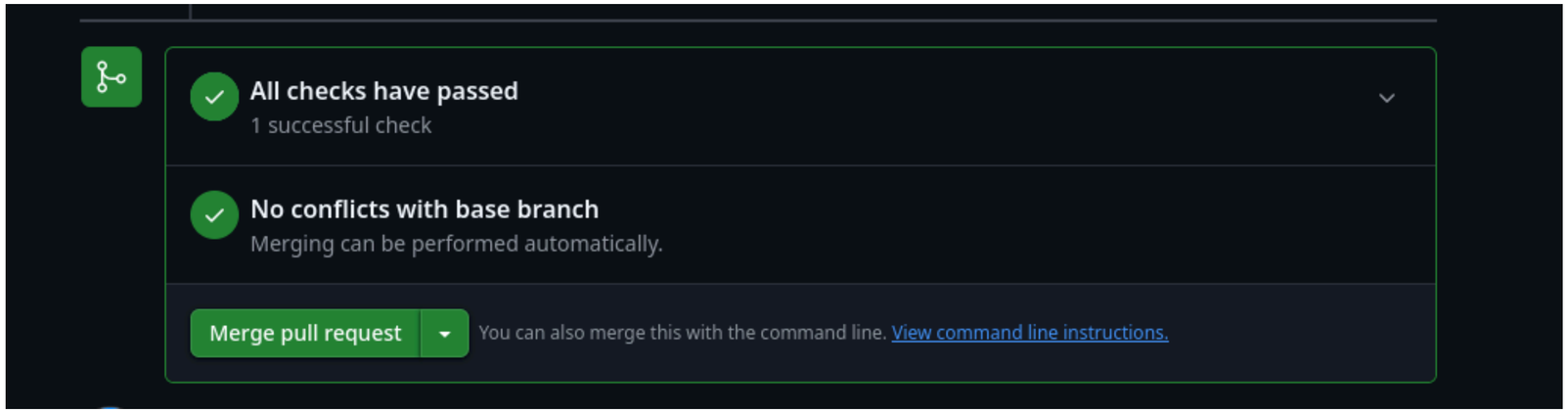
Action Strategy powered by W&B

GitHub Actions automate our CI/CD pipelines. They connect with Weights & Biases (W&B) to register new model versions, track experiments, and promote proven models to production via the W&B Registry.

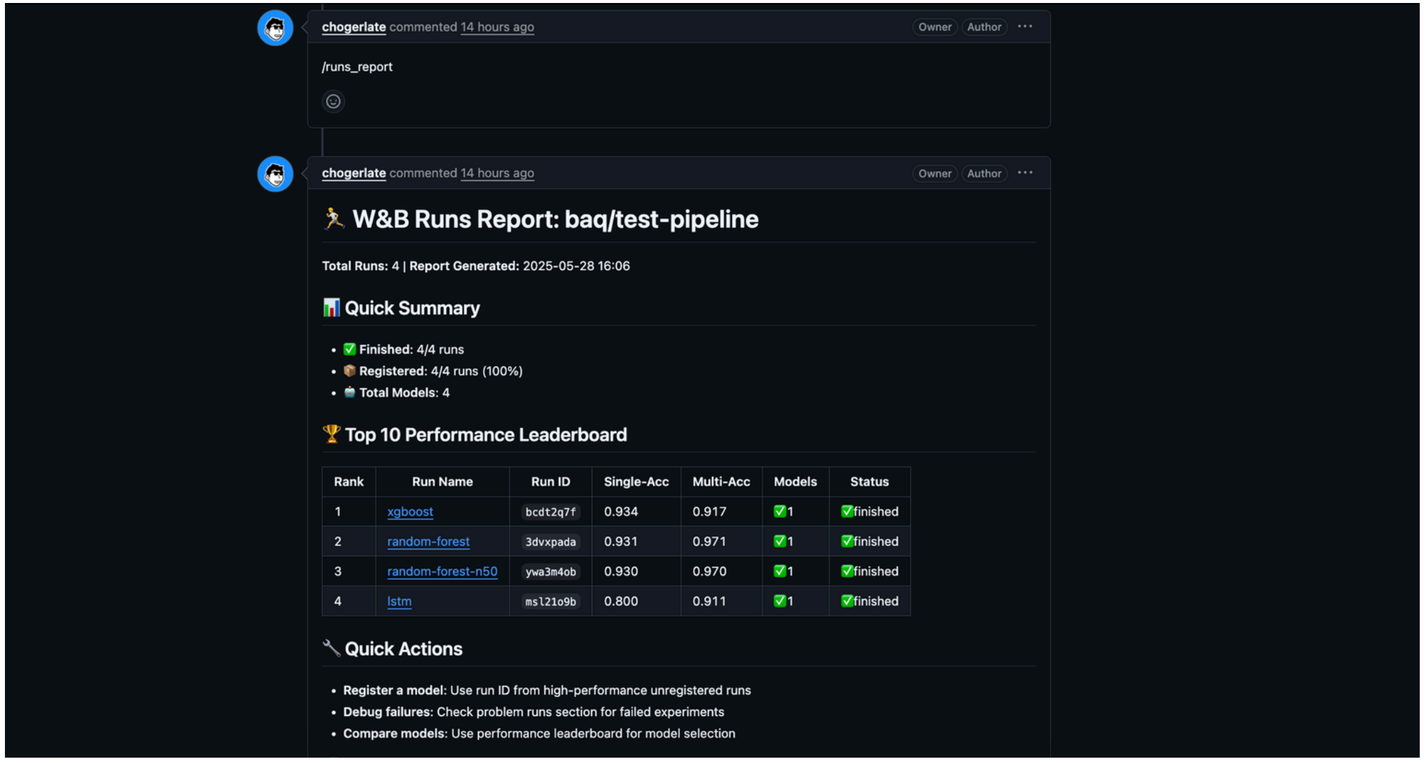


The project integrates several custom GitHub Actions to streamline experiment tracking, model registration, and deployment workflows. These actions interact directly with the Weights & Biases (WandB) platform and are triggered through pull request comments on GitHub.

- **Dependencies Checking** : Automatically checks project dependencies (e.g., from requirements.txt) during pull requests to ensure all packages are valid and installable.



- **/runs_report** : Fetches and displays a summary of recent WandB experiment runs, including key metrics and hyperparameters, as a comment in the pull request.tates easy comparison and selection of model candidates during code review.



chogerlate commented 14 hours ago — Owner Author

/runs_report

chogerlate commented 14 hours ago — Owner Author

W&B Runs Report: baq/test-pipeline

Total Runs: 4 | **Report Generated:** 2025-05-28 16:06

Quick Summary

- ✅ **Finished:** 4/4 runs
- 📦 **Registered:** 4/4 runs (100%)
- 🤖 **Total Models:** 4

Top 10 Performance Leaderboard

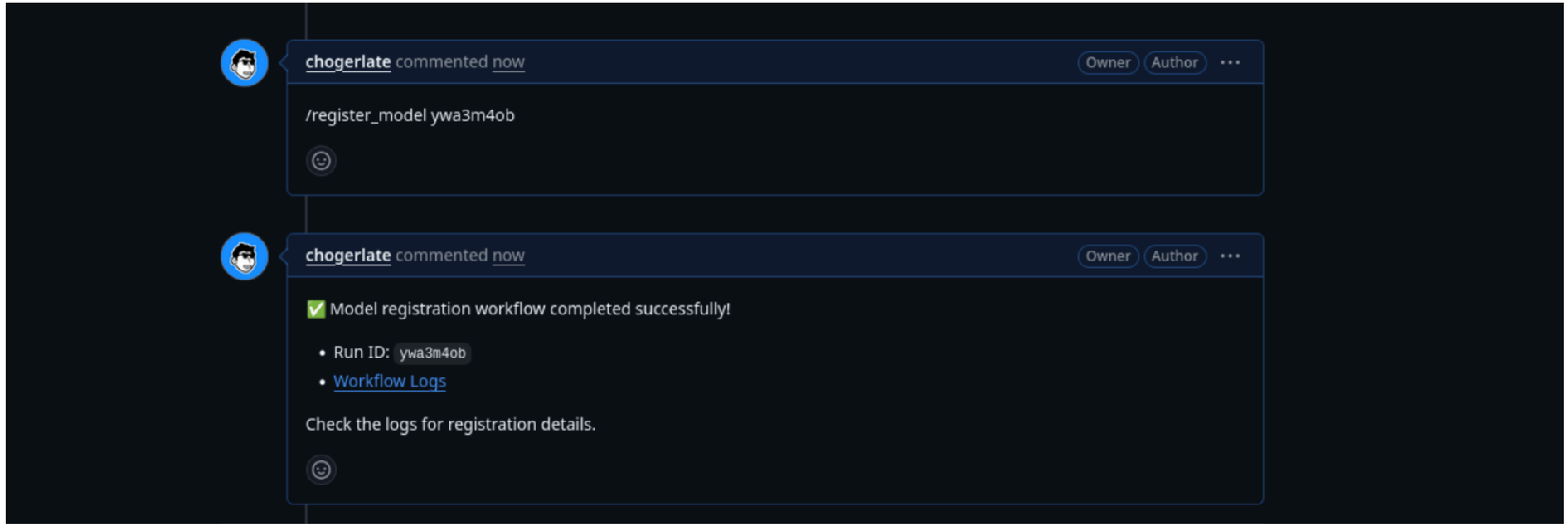
Rank	Run Name	Run ID	Single-Acc	Multi-Acc	Models	Status
1	xgboost	bcdt2q7f	0.934	0.917	✅1	✅finished
2	random-forest	3dvxpada	0.931	0.971	✅1	✅finished
3	random-forest-n50	ywa3m4ob	0.930	0.970	✅1	✅finished
4	lstm	msl21o9b	0.800	0.911	✅1	✅finished

Quick Actions

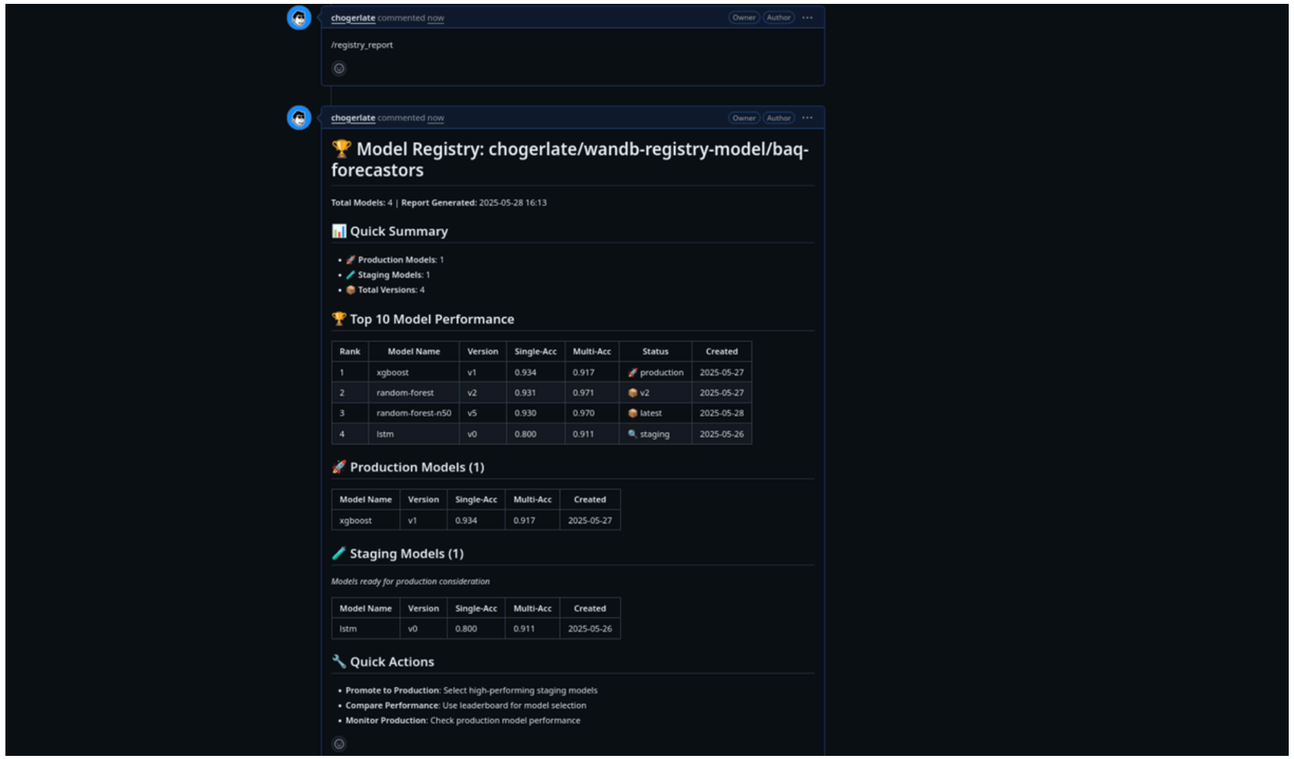
- **Register a model:** Use run ID from high-performance unregistered runs
- **Debug failures:** Check problem runs section for failed experiments
- **Compare models:** Use performance leaderboard for model selection

1 Review Runs → 2 Register Candidates → 3 Compare Models → 4 Promote Best Model

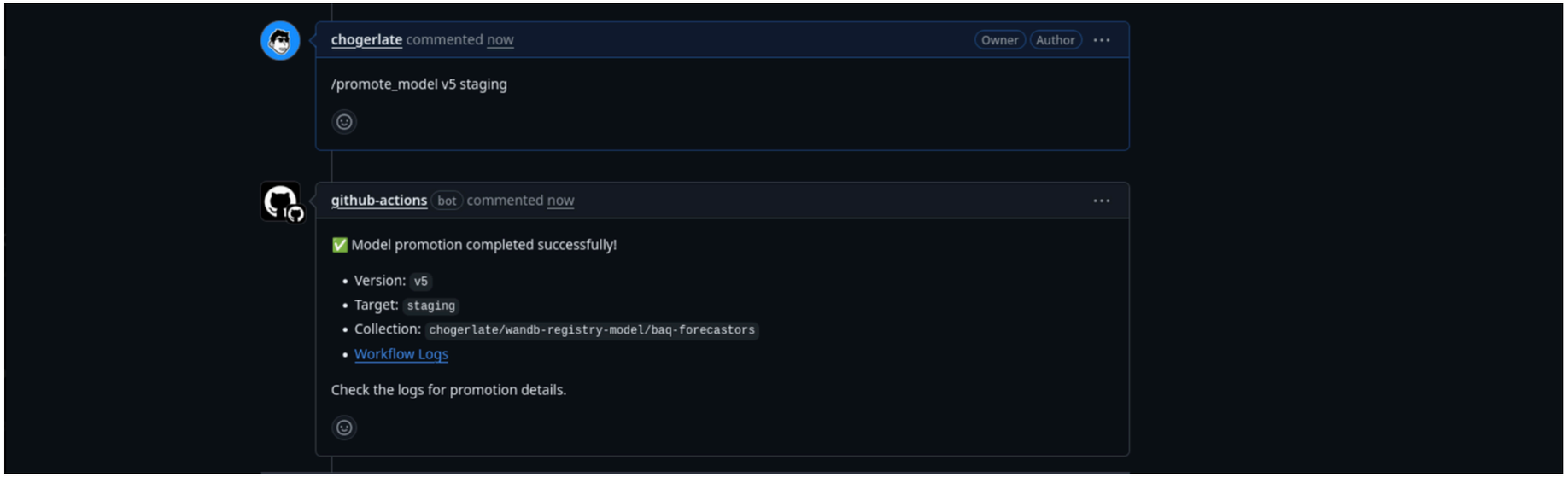
- **/register_model {run_id} :** Registers a trained model from the specified WandB run_id into the WandB model registry and confirms the registration in the pull request.



- **/registry_report :** Lists all registered models in the WandB registry, showing names, versions, and aliases in the pull request for easy reference.

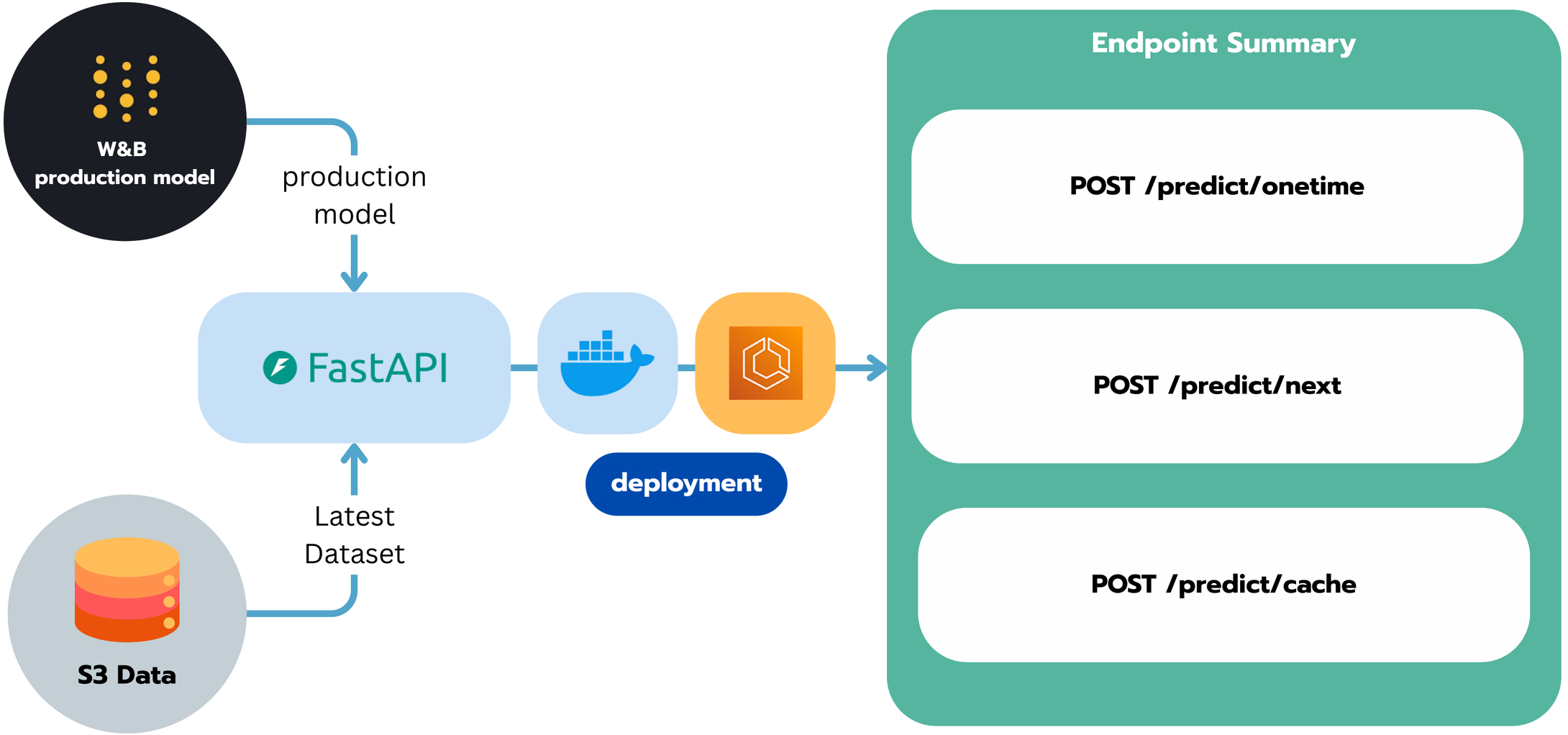


- **/promote_model {version} {target_alias} :** Promotes a registered model version by assigning it a new alias (e.g., production), and confirms the update in the pull request.

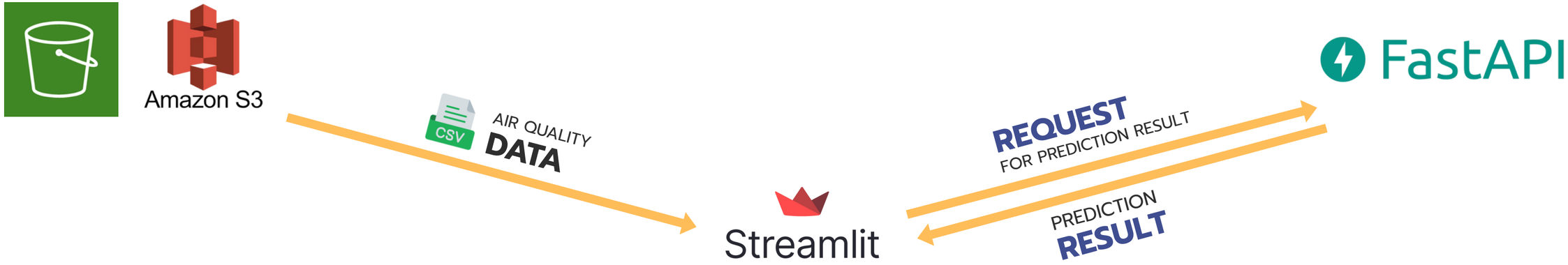


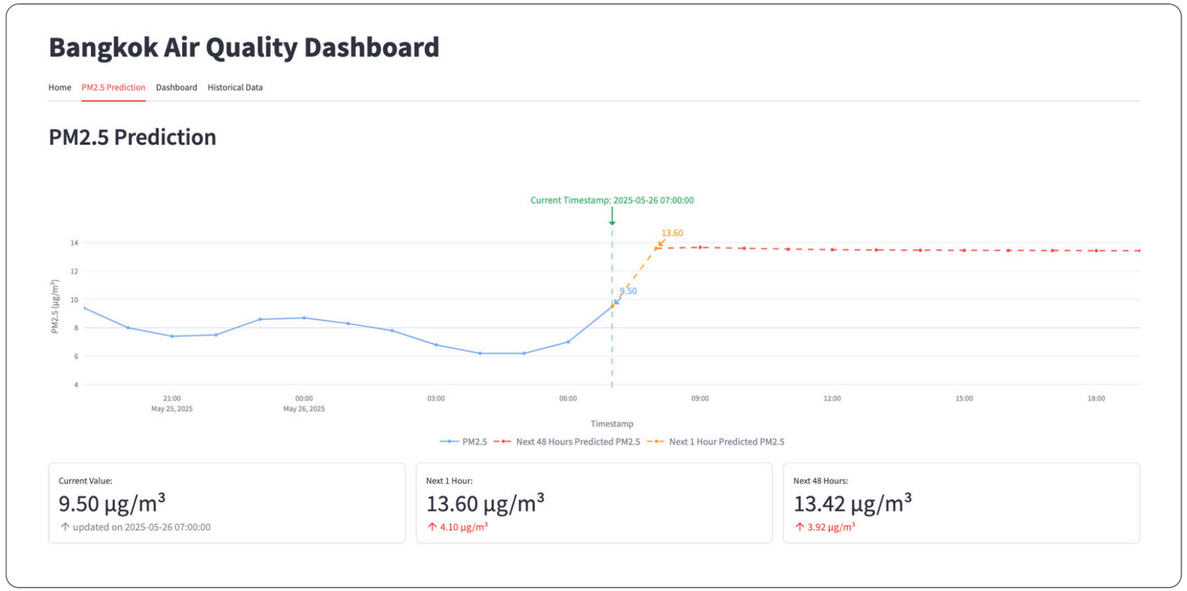
1 Review Runs → 2 Register Candidates → 3 Compare Models → 4 Promote Best Model

FASTAPI - APPLICATION BACKEND



STREAMLIT-FRONTEND

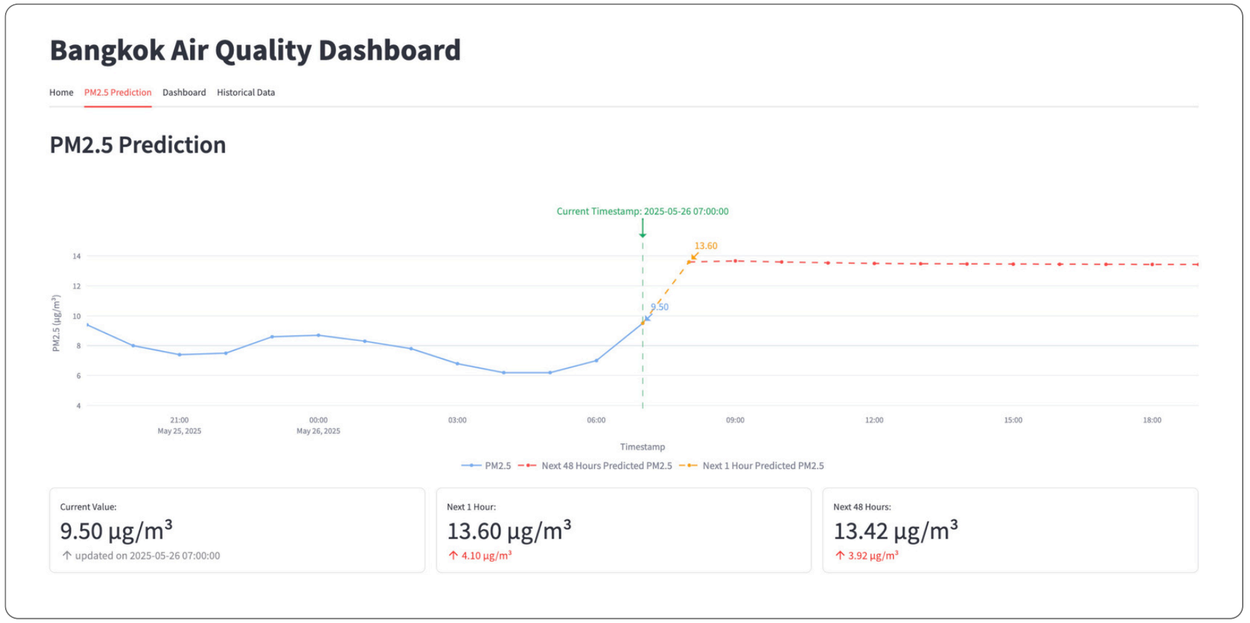




STREAMLIT

DEPLOYMENT

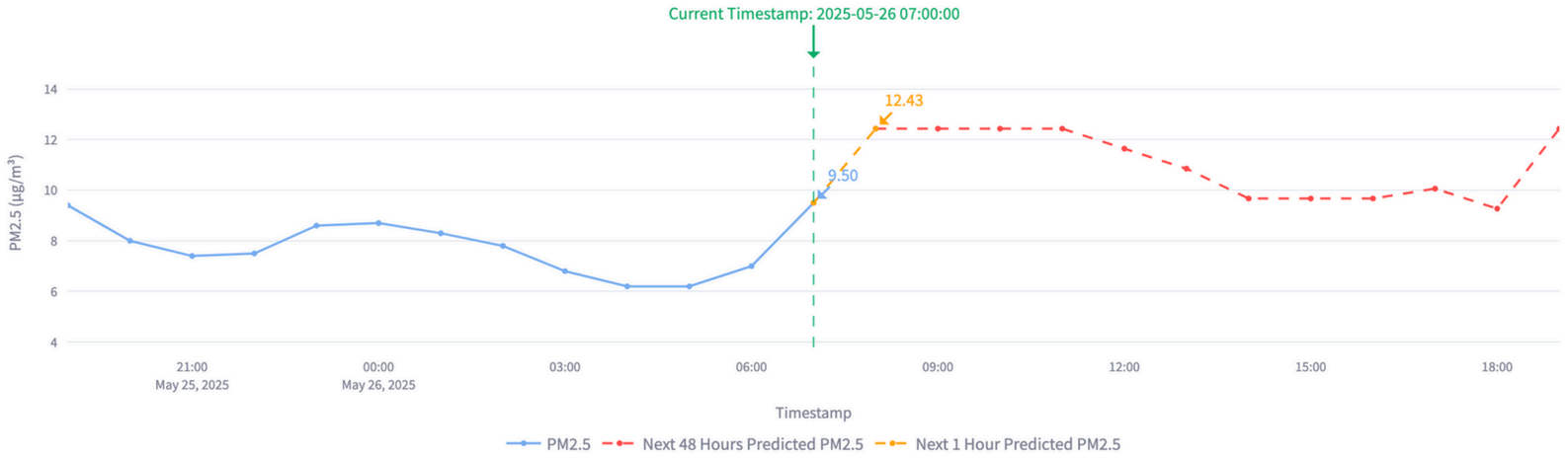
WEB LINK: *https://baq-frontend.streamlit.app/*



Bangkok Air Quality Dashboard

Home | PM2.5 Prediction | Dashboard | Historical Data

PM2.5 Prediction



Current Value:	Next 1 Hour:	Next 48 Hours:
9.50 µg/m³	12.43 µg/m³	11.64 µg/m³
↑ updated on 2025-05-26 07:00:00	↑ 2.93 µg/m³	↑ 2.14 µg/m³

Bangkok Air Quality Dashboard

Home PM2.5 Prediction Dashboard Historical Data

Dashboard Insights

Select Date Range

2023/05/28 – 2025/05/26

📌 Key Performance Indicators

Average PM2.5	Average PM10	Average Temperature	Average UV Index
25.8 µg/m³	35 µg/m³	28.6 °C	2.0

📈 Trends Over Time

PM2.5 & PM10 Over Time

Bangkok Air Quality Dashboard

Home | PM2.5 Prediction | Dashboard | Historical Data

Historical Data

Historical Data Table

	time	temperature_2m_(°C)	relative_humidity_2m_(%)	dew_point_2m_(°C)	apparent_temperature_(°C)	precipitation_(mm)	rain_(mm)	weather_code_(wmo_code)	pressure_msl_(hPa)	s
0	2023-05-28 07:00:00	27.476	83.4794	24.426	32.5511	0	0	3	1006	
1	2023-05-28 08:00:00	27.376	83.4682	24.326	32.5483	0.1	0.1	51	1007.4	
2	2023-05-28 09:00:00	28.326	78.7292	24.276	33.8897	0.1	0.1	51	1007.9	
3	2023-05-28 10:00:00	29.576	72.5763	24.126	35.3263	0	0	3	1007.7	
4	2023-05-28 11:00:00	31.526	63.1803	23.676	37.2161	0	0	3	1006.9	
5	2023-05-28 12:00:00	32.576	58.1156	23.276	38.4703	0	0	3	1006.3	
6	2023-05-28 13:00:00	33.326	55.8839	23.326	39.2086	0	0	3	1005.1	
7	2023-05-28 14:00:00	33.826	55.167	23.576	39.5884	0	0	2	1004.1	
8	2023-05-28 15:00:00	34.326	52.8467	23.326	40.3108	0	0	2	1003.3	
9	2023-05-28 16:00:00	30.876	68.5833	24.426	35.9207	2.8	2.8	63	1003.7	

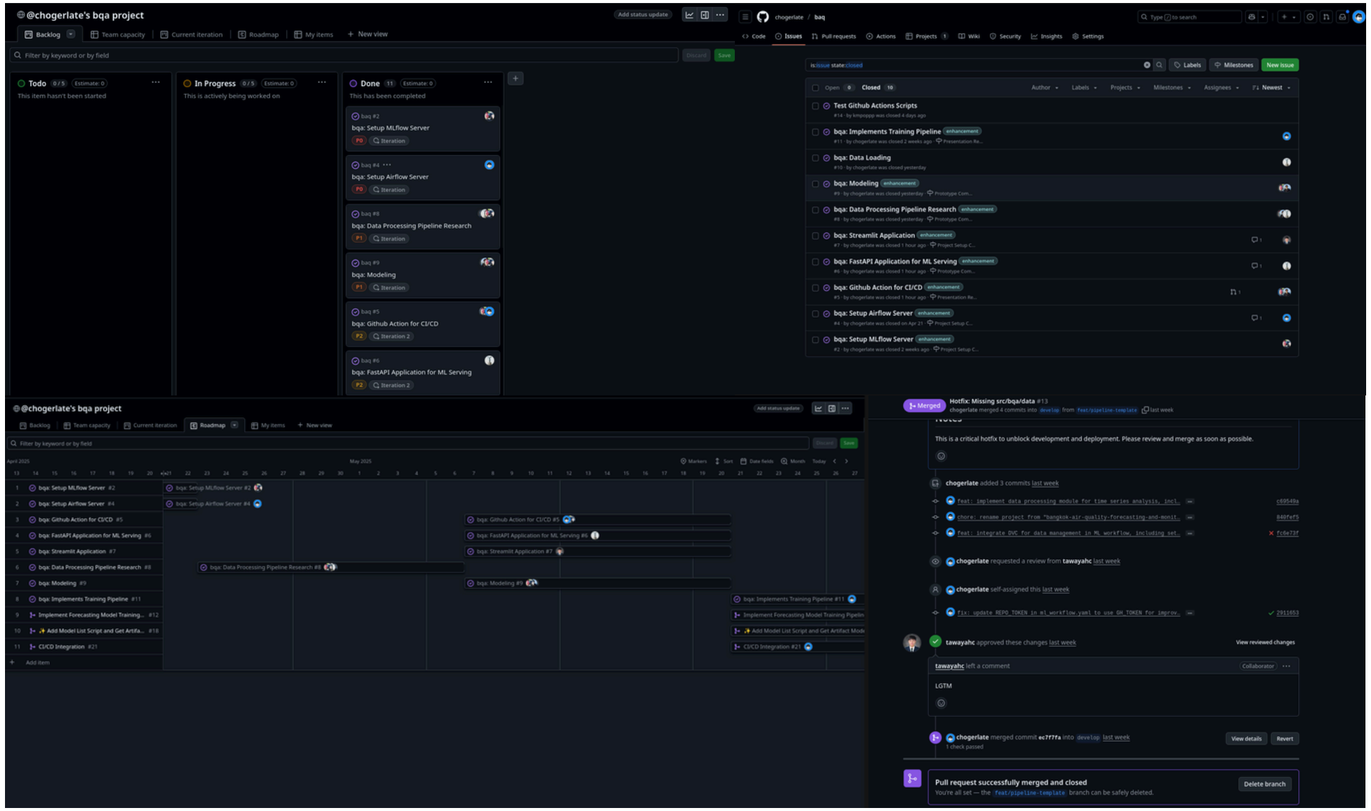
Historical Data Chart

📌 Select columns to plot:

temperature_2… ×

GITHUB PROJECT

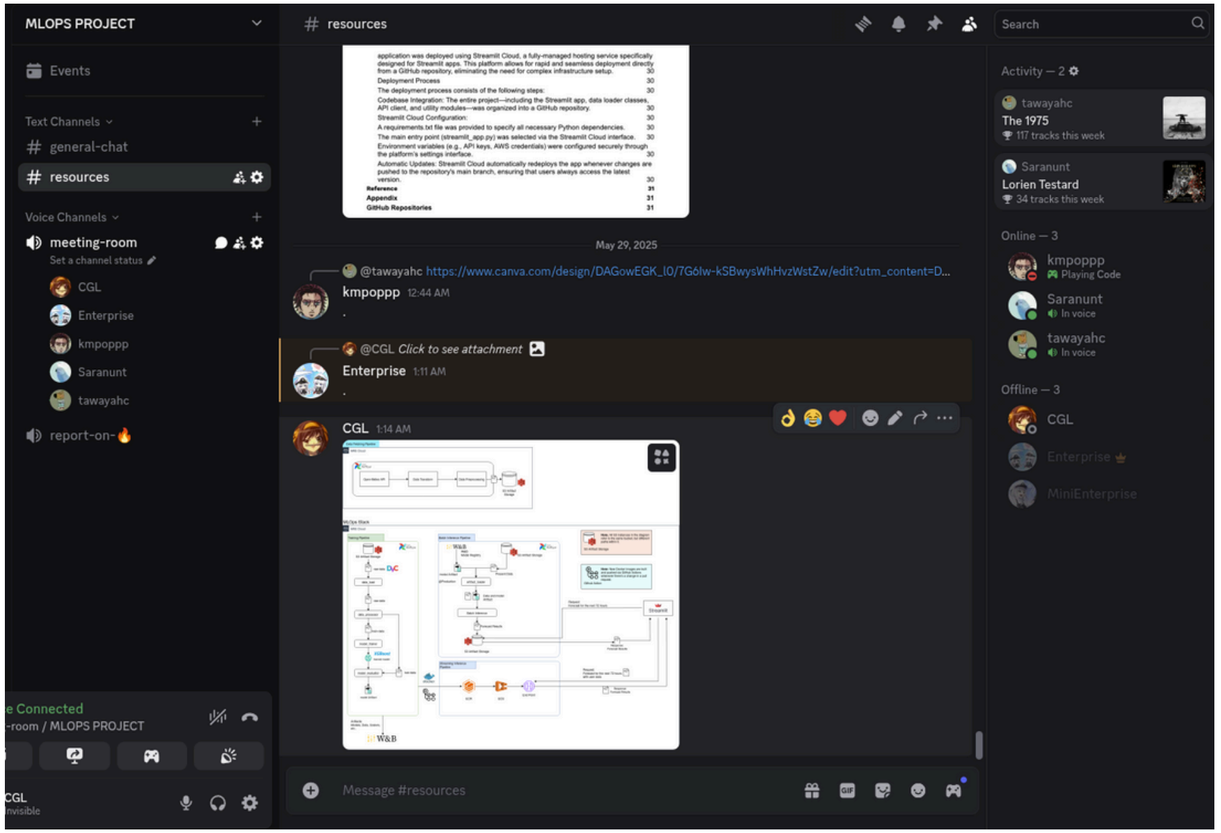
We use GitHub Projects for visual tracking of tasks and Issues to define work, assign owners, and set deadlines. This keeps our team aligned and our project timeline on track.



Github Management Workflow

DISCORD

We use Discord as our main messenger app for real-time communication, quick discussions, and sharing updates. It keeps our team connected and informed instantly.



Our Discord Server

OUR TEAM

Members

Student ID	Name-Surname	Responsibilities
65070501012	Chayawat Anaroch	Data Scientist, MLOps Engineer
65070501052	Siwarat Laoprom	ML Engineer, MLOps Engineer
65070501065	Kamolpop Poonsawat	Project Manager, Data Scientist
65070501084	Poomrapee Moungnoi	Data Engineer, Data Scientist
65070501092	Bhagya Saranunt	Data Scientist, ML Engineer