

# Project DSI321: Real-Time Data Pipeline with Visualization

---

## Bangkok Air Quality Dashboard with Streamlit, Prefect & LakeFS

 Created at Last commit today

---

### Overview

This project, built for **DSI321: Big Data Infrastructure**, monitors **Bangkok's air quality in near real-time** using hourly PM2.5 and AQI data from Air4Thai. It also **forecasts pollution 6 hours ahead** with ARIMA, helping support public health and city planning.

All components run in Docker for consistency and reproducibility.

---

### Tech Stack Summary

Layer	Tools
Orchestration	Prefect
Containerization	Docker
Data Versioning	LakeFS
Visualization	Streamlit
Forecasting Model	ARIMA
Notebook IDE	JupyterLab
Data Source	<a href="#">Air4Thai PM2.5 API</a>

### Key Features

- Live AQI/PM2.5 updates from Bangkok stations
  - 6-hour forecasting per station using ARIMA
  - Interactive dashboard with maps, charts, and tables
  - Automated flows using Prefect
  - Versioned data with LakeFS
  - Fully containerized with Docker
- 

### Data Schema (processed for dashboard)

Column	Type	Description
timestamp	datetime	Measurement time

Column	Type	Description
stationID	string	Station code
nameTH	string	Station name (Thai)
areaTH	string	Area name (Thai)
district	string	Bangkok district name
lat, long	float	Coordinates
AQI.aqi	int	Air Quality Index (0–500)
PM25.value	float	PM2.5 ( $\mu\text{g}/\text{m}^3$ )

## Data Quality Checks

- 1,000+ records across stations
- 24+ hourly records per station
- $\geq 90\%$  completeness
- No duplicated rows or object-type columns

 [View full QA notebook](#)

## Quick Start

### 1. Clone the Repo

```
git clone https://github.com/Nxmfqn/dsi321_2025.git  
cd dsi321_2025
```

### 2. Start All Services

```
docker-compose up --build -d
```

### 3. Access Local Services

Service	URL
LakeFS	<a href="http://localhost:8001">http://localhost:8001</a>
JupyterLab	<a href="http://localhost:8888">http://localhost:8888</a>
Prefect	<a href="http://localhost:4200">http://localhost:4200</a>
Streamlit	<a href="http://localhost:8502">http://localhost:8502</a>

💡 Login for LakeFS: `access_key / secret_key`

Then, create the LakeFS repo:

```
lakectl repo create lakefs://dust-concentration
```

## 📝 Upload & Forecast (First-Time Setup)

Upload Initial Data

```
docker exec -it dsi321-jupyter-1 bash  
python upload.py
```

Run Forecast Scripts

```
python pipeline/getdata.py  
python pipeline/forecasting.py
```

or trigger via **Prefect UI** at <http://localhost:4200>

## ⌚ Automate with Prefect

Schedule flows to run every hour:

```
# Ingest data (runs at :25)  
python deploy.py  
  
# Forecast data (runs at :27)  
python deploy_ml.py
```

## 📊 Dashboard Overview

Deploy ⚙️

### รายงานคุณภาพอากาศในกรุงเทพมหานคร

🕒 ข้อมูลล่าสุดเมื่อ: 18/05/2025 23:47:29

ค้นหาสถานที่หรือเขต

สำนักงานเขตคลองเตย (คลองเตย)

## ສ້ານັກງານເບຕຄລອງເຕຍ (ຄລອງເຕຍ)

**AQI 38 – Good**

PM2.5 – 19.9  $\mu\text{g}/\text{m}^3$

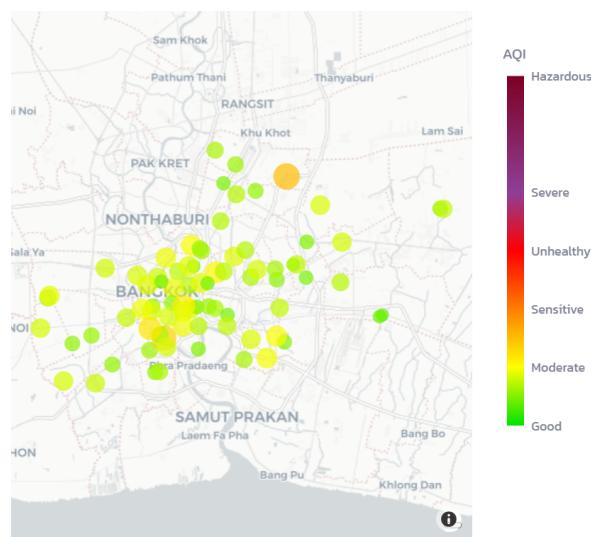
### Dashboard

ຄ່າເຈລ່ຍຄຸນພາພອາກາສກາຍໃນກຽງເຖວຍ

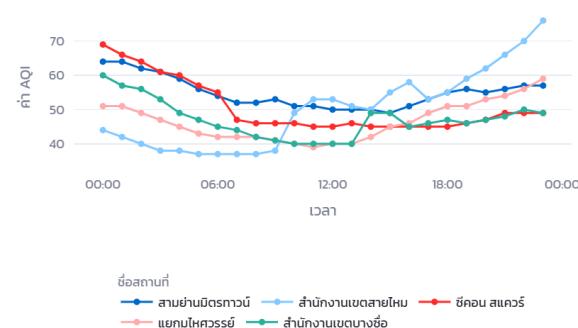
ຄ່າເຈລ່ຍ AQI  
**AQI 37 – Good**

ຄ່າເຈລ່ຍ PM2.5  
**19.5  $\mu\text{g}/\text{m}^3$**

### ແພນທີ່ຄຸນພາພອາກາສ



### 5 ສະຖານທີ່ໃນກຽງເຖວຍທີ່ມີຄ່າ AQI ສູງທີ່ສຸດ (today)



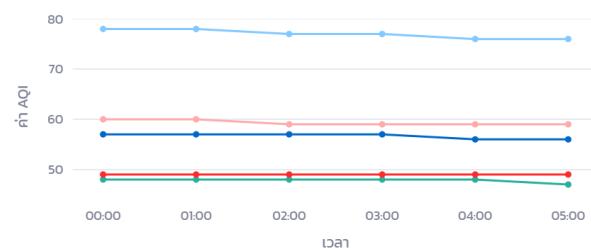
### ພຍາກຮັນຄຸນພາພອາກາສລ່ວງໜ້າ

ເລືອດສະບັບທີ່

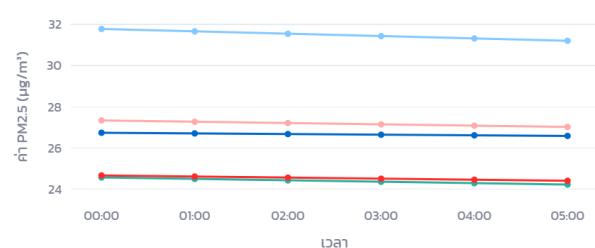
ສ້ານັກງານເບຕສາຍໄ... x ແຍກນໄທຂວຽຮ x ສະບັບມືຕຣາວນີ x ຫັກອນ ສເຄວົກ x ສ້ານັກງານເບຕບານ...

x v

### ພຍາກຮັນ AQI



### ພຍາກຮັນ PM2.5



ສະຖານທີ່  
— ● ສະບັບມືຕຣາວນີ — ● ສ້ານັກງານເບຕບານຊົ້ວ — ● ສ້ານັກງານເບຕສາຍໄຫ້  
— ● ແຍກນໄທຂວຽຮ — ● ຫັກອນ ສເຄວົກ

ສະຖານທີ່  
— ● ສະບັບມືຕຣາວນີ — ● ສ້ານັກງານເບຕບານຊົ້ວ — ● ສ້ານັກງານເບຕສາຍໄຫ້  
— ● ແຍກນໄທຂວຽຮ — ● ຫັກອນ ສເຄວົກ

### ຂໍ້ມູນກັ້ນໜົດ (ຊັ້ນໂມງສໍາສັດ)

timestamp	nameTH	district	AQI.aqi	PM25.value
18/05/2025 23:00	ນະວັທະຍາສໍາຮັບກົງປານສົ່ງເຈົ້າພະຍາ	ຮັບອີ	29	16.5
18/05/2025 23:00	ຮັນນັນທາງທະວານນາຍເລີນ 3902	ນາງຂົນເທິບຍນ	40	20.9
18/05/2025 23:00	ກາຣເຄຫະໝູນນໍາໄວໜ້າງ	ດິນແດງ	53	25.5
18/05/2025 23:00	ໂຮງເຮັດວຽກທະວາງ	ບານນາງວາ	36	19.1
18/05/2025 23:00	ໂຮງພານາຄຈຸກ້າລົງກຣົມ	ປັກຫຼວນ	43	22.2
18/05/2025 23:00	ກາຣໄຟຟ້າມອຍຮັບອີ	ຮັບອີ	43	22.1
18/05/2025 23:00	ສການີ່ຕ່າງຈຳນວຍຄວາມລົ້ອມຂັ້ນ	ວັງທອນຫລາງ	44	22.5

18/05/2025 23:00	การค่าหะนุชตินดอง	ตันแಡง	45	23.1
18/05/2025 23:00	กรมประชาสัมพันธ์	พญาไท	22	13.4
18/05/2025 23:00	โรงเรียนคินทร์เดชา (สิงห์ สิงหเสนี)	วังทองหลาง	34	18.5
18/05/2025 23:00	สำนักงานเขตคลองสานฯ	เมืองกุน	21	12.4
18/05/2025 23:00	สำนักงานเขตคลองสานฯ	คลองสามวา	43	22.3
18/05/2025 23:00	สำนักงานเขตคลองหลอม	คลองหลอม	32	17.6

- 🔍 Select stations and view real-time AQI/PM2.5
- 📈 Line charts for top polluted stations
- 🌐 AQI map of Bangkok with color-coded bubbles
- 📊 Forecast and data tables for all stations

## ⌚ Forecasting Details

- **Model:** ARIMA(1,0,1) per station
- **Target:** PM2.5 (float) & AQI (int)
- **Stored at:** `lakefs://dust-concentration/main/forecast/forecast.parquet`
- **Forecast horizon:** Next 6 hours
- **Excludes:** Incomplete or flat-line stations

## 📁 Repo Structure

```

.
├── data/
│   ├── data.parquet/year=2025/month=5/day=XX/hour=XX/
│   ├── SCHEMA.md
│   └── check_data_quality.ipynb
├── img/
│   └── dashboard_demo.png
└── pipeline/
    ├── bangkok_districts.geojson
    ├── getdata.py
    ├── forecasting.py
    ├── deploy.py
    ├── deploy_ml.py
    └── upload.py
└── prefect/
    ├── Dockerfile.jupyter
    ├── Dockerfile.prefect-worker
    ├── requirements.txt
    └── wait-for-server.sh
└── visualization/
    ├── .streamlit/config.toml
    └── app.py
└── .gitignore

```

```
├── README.md  
└── docker-compose.yml
```