Final Presentation

Air Quality Trends and Thermal Power Correlation in Korea (2003–2024)

Team 1
21102061 Hwang Hyunmin
21102052 Lee Jeongyun
23102020 Lee Sodam
23102025 Lee Haneol

https://github.com/han3o1/BDP-Airquality-Analysis.git

Contents

Problem Definition

Problem

Sustained air quality degradation since 2003, driven by economic growth and increased energy demand.

Hypothesis:
The rise in national thermal power generation has negatively impacted air quality.

Sustained air quality degradation since 2003,
driven by economic growth
and increased energy demand.

Additionally, all essential preprocessing steps:
data collection → cleaning
→ storage → monthly aggregation
→ normalization → analysis
were performed manually, which resulted in reduced data consistency and reproducibility.

Problem Definition

Goal

- Quantitatively analyze long-term trends linking energy demand and air quality.
- Systematically explore the correlation between thermal power output and pollutant concentration.

Establishment of a "Self-updating Analytical Pipeline" by automating data collection, loading, and analysis on a monthly basis.

- Based on the analysis, we can predict the future relations between thermal power and air quality.
- Display the overall relation in web based dashboard for better insight.

Dataset

0.2 데이터 생성주기

※ 에어코리아 OpenAPI 서비스 내 오퍼레이션 데이터 생성주기

API 명(국문)	상세기능명(국문)	상세기능명(영문)	데이터 생성주기
	측정소별 실시간 측정정보 조회	getMsrstnAcctoRltmMesureDnsty	매시 15 분 내외

region	station_code	station_name	date_time	SO2	CO	O3	NO2	PM10	PM25	address
서울 중구	111121	중구	2.004E+9	0.001	0.7	0.038	0.008	35		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.001	0.7	0.038	0.008	35		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.001	0.7	0.04	0.007	33		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.001	0.7	0.036	0.01	27		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.001	0.8	0.027	0.019	30		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.001	0.8	0.013	0.04	28		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.002	0.9	0.009	0.045	35		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.003	1	0.009	0.048	41		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.004	0.9	0.013	0.044	45		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.004	0.9	0.021	0.036	56		서울 중구 덕수궁길 15
서울 중구	111121	중구	2.004E+9	0.003	0.8	0.03	0.026	47		서울 중구 덕수궁길 15

Meta	측정소코드 → **station_code**
	측정소명 → **station_name**
Time	측정일시 → **date_time**
Pollutants	아황산가스→ **SO2**
	미세먼지 → **PM10**
	초미세먼지 → **PM25**
	(기타) → **NO2, O3, CO**

Air Korea Data

- **Iterative Collection (Region Looping):**

Since the API does not support a nationwide bulk download, the system iterates through a list of **17 administrative divisions** (e.g., Seoul, Busan, Jeju) to fetch data sequentially.

- **Version Control (ver=1.5):**

Utilized the ver=1.5 parameter to retrieve the most granular data schema, including **PM2.5** and detailed station metadata.

- **Schema Normalization:**

Automatically maps Korean JSON keys (e.g., 미세먼지농도) to English column names (e.g., PM10) for compatibility with the analytics engine (Spark/Hive).

Dataset

	2003	2004	2005	2006
1	13,222,237	12,648,481	14,089,063	15,062,281
2	12,097,968	11,971,217	12,389,177	12,766,622
3	11,536,474	12,516,914	12,754,514	13,075,849
4	10,576,434	11,180,184	11,938,918	11,337,712
5	11,114,352	11,333,706	11,871,907	11,298,897

Category	Column Name	Description
Time	date_time	Raw: Hourly → Processed: Monthly
Category	fuel_type	Filtered for fossil fuels only
Measure	power_value	Power Trading Volume (MWh)

Thermal Power

- **Dynamic Pagination:**

Implemented While-Loop logic to perform a full scan of millions of annual rows, preventing data loss due to API page limits.

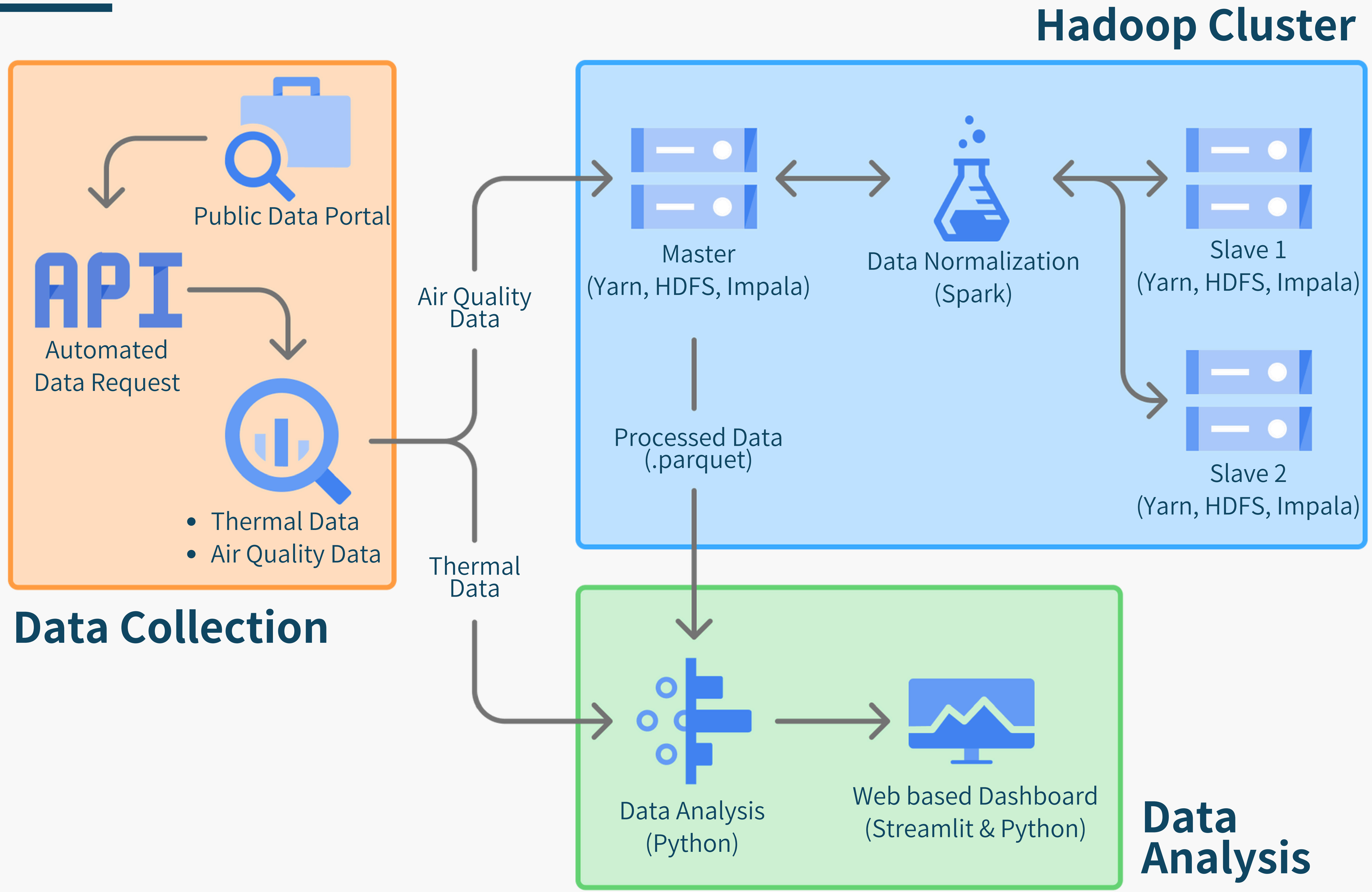
- **Smart Filtering:**

Extracted only fossil fuel sources linked to air pollution, excluding irrelevant sources like Nuclear or Solar power.

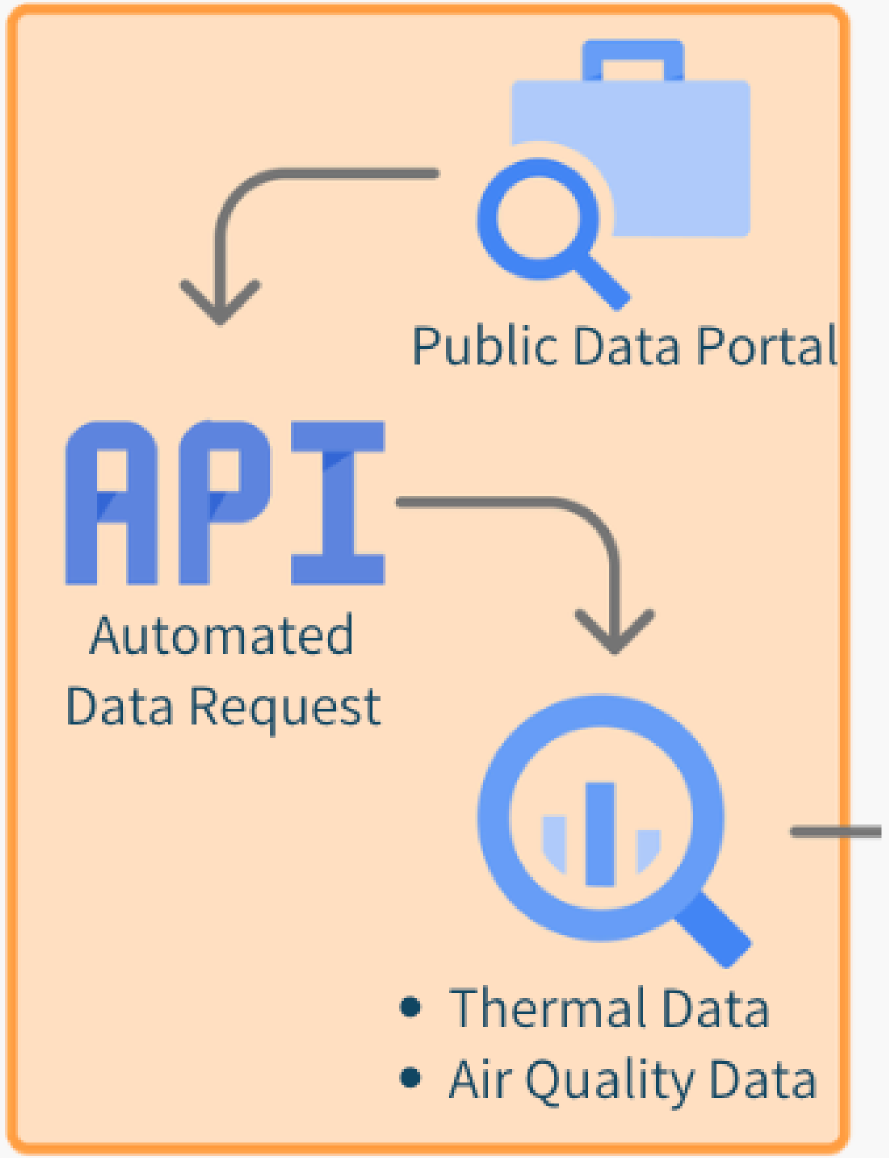
- **Time-based Aggregation (Hourly → Monthly):**

Unlike simple downloading, the system aggregates raw hourly data into monthly sums during the collection phase to align with the analysis timeframe.

System Architecture



System Architecture - Data Collection

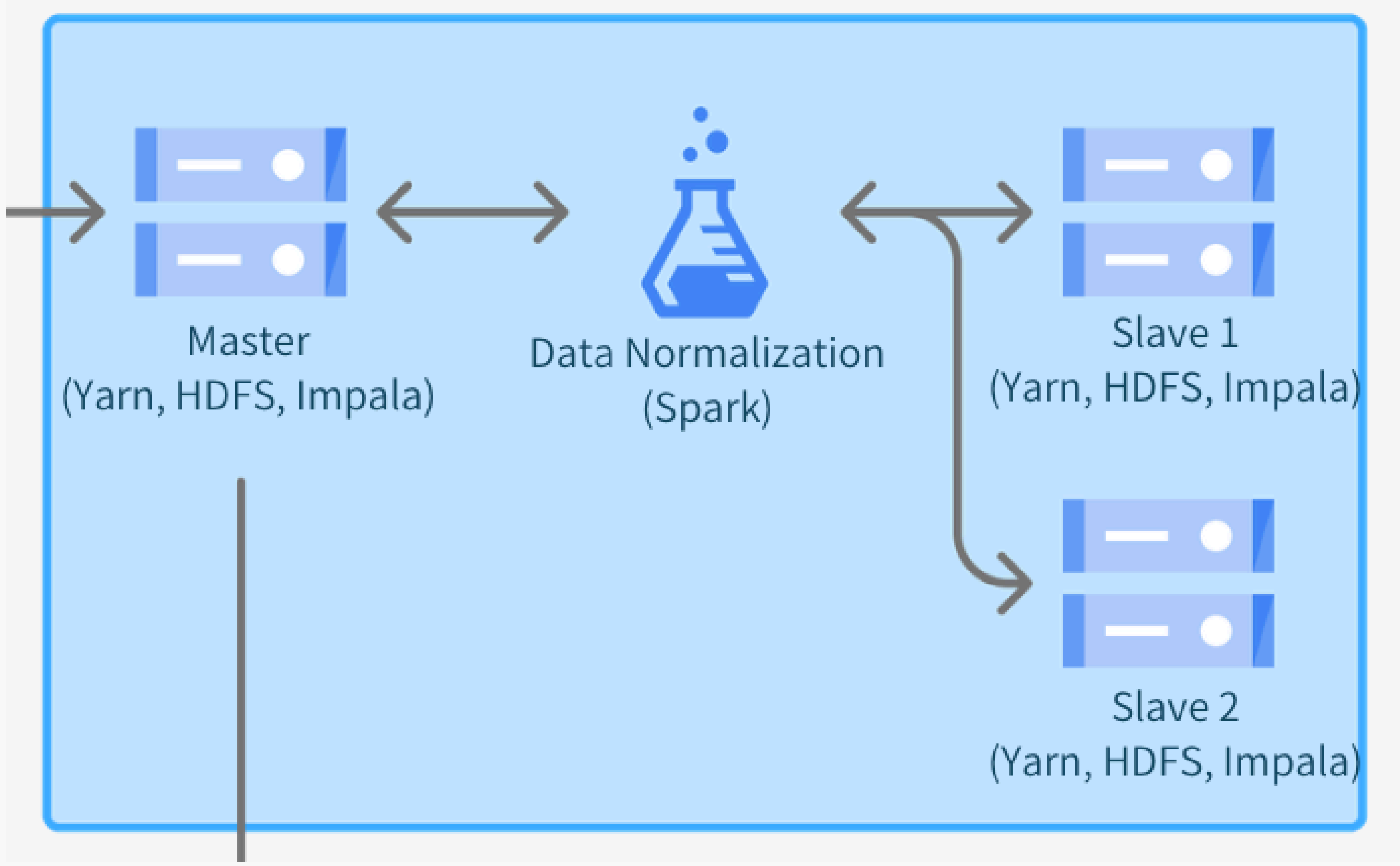


- **Source**: Official Government Portal (data.go.kr) for data reliability.
- **APIs**: Integrated AirKorea (Air Quality) & KEPCO (Power Trading Volume) Open APIs.

- **Python Engine:** Implemented dynamic pagination (handling 1M+ rows) and schema normalization (unifying column names).
- **Scheduler:** Linux Cron triggers script monthly for zero-maintenance updates.

- **Thermal Power:** Specifically filtered for **fossil fuels (Coal, LNG)** and aggregated **hourly data** into **monthly statistics.**
- **Air Quality:** Secured 22-year time series (2003–2024) for major pollutants (SO2, NO2, PM10, etc.)

System Architecture - Hadoop Cluster



Cluster Nodes and Storage

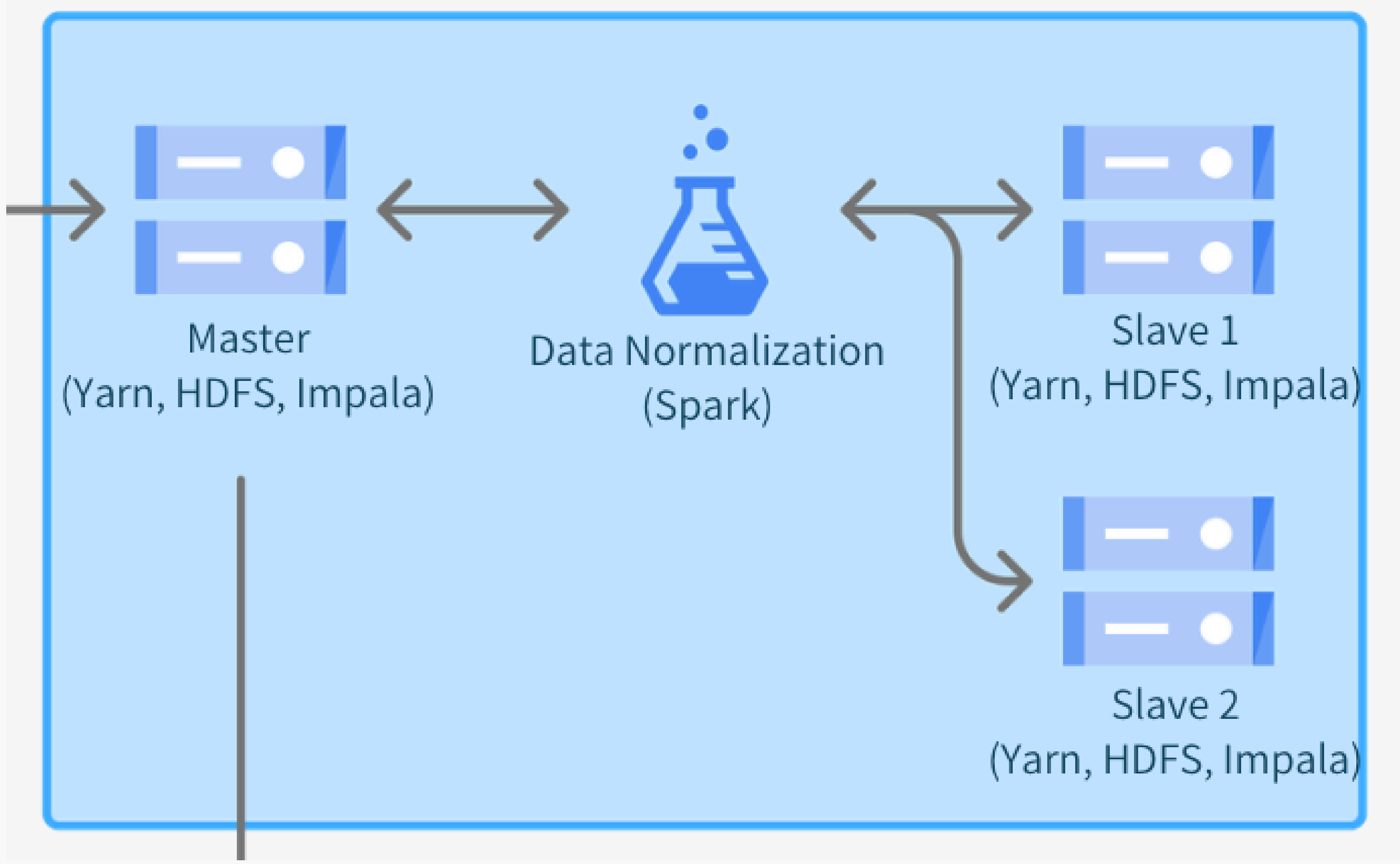
Hadoop Cluster Nodes

- **1 Master**: NameNode, ResourceManager, Metastore
- **2 Slaves**: Datanodes
  - → ~70% faster data loading and Spark processing
- **Data replication factor**: 2

Data Storage Details

- Connected with **Local Network & Sync hostname**
- File transfer to VM local storage via VMWare Shared Folder
- **HDFS Partitioning** → /year=YYYY/month=MM/YYYY-MM.csv
- **Additional Disk added in Master Node** for airqualithy data storage

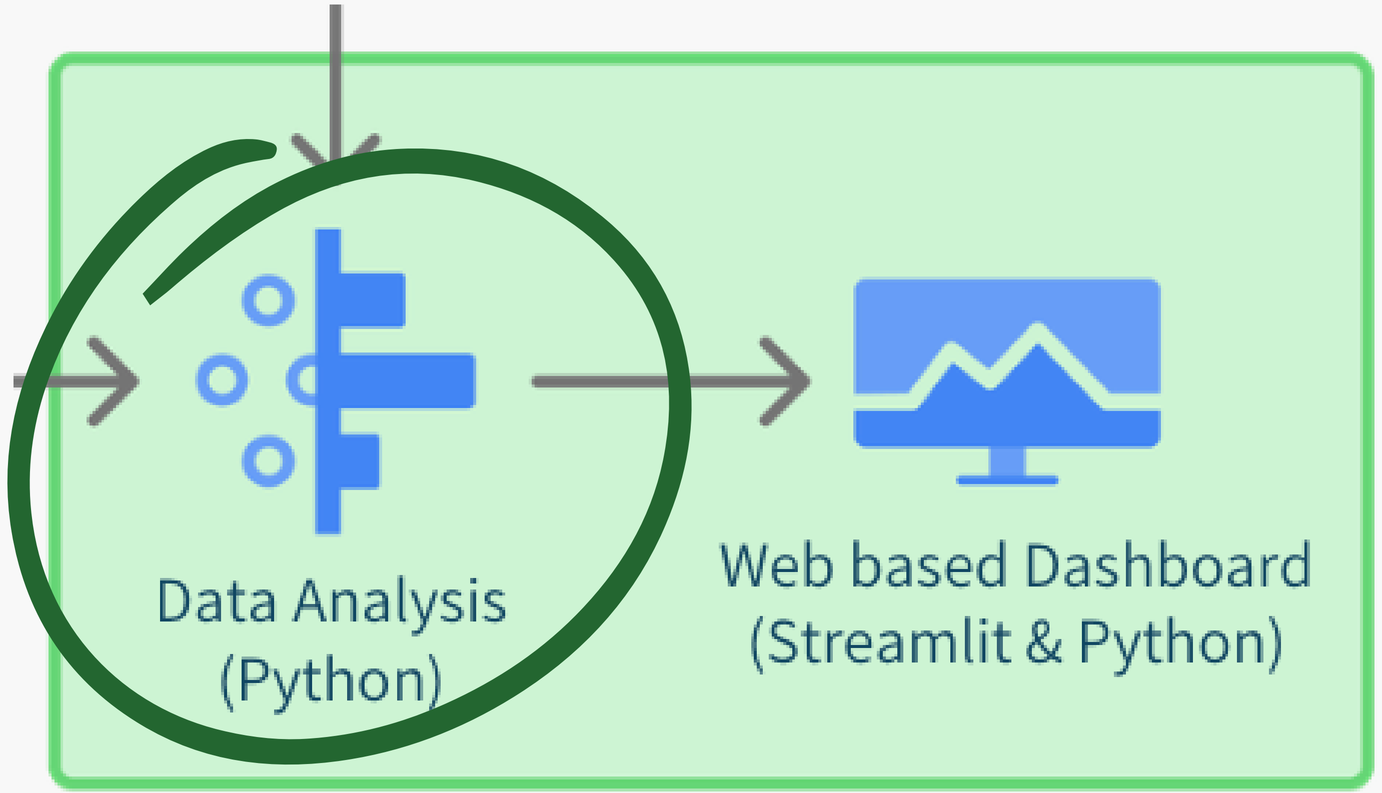
System Architecture - Hadoop Cluster



Data Preprocessing & Normalization

- **CSV Parser Implementation**
  - read file as textfile RDD
  - extract 1st row as table schema
  - for each rows split by comma → CSV parsing
- **Missing value** → mean subsititution
- **Outlier handling** → Z-score Normalization
- **Parquet Transformation** (partitioning kept, /year=YYYY/month=MM/YYYY-MM.parquet)
- Quick access into processed data through **Impala (monthly data aggregation)**

System Architecture - Data Analysis



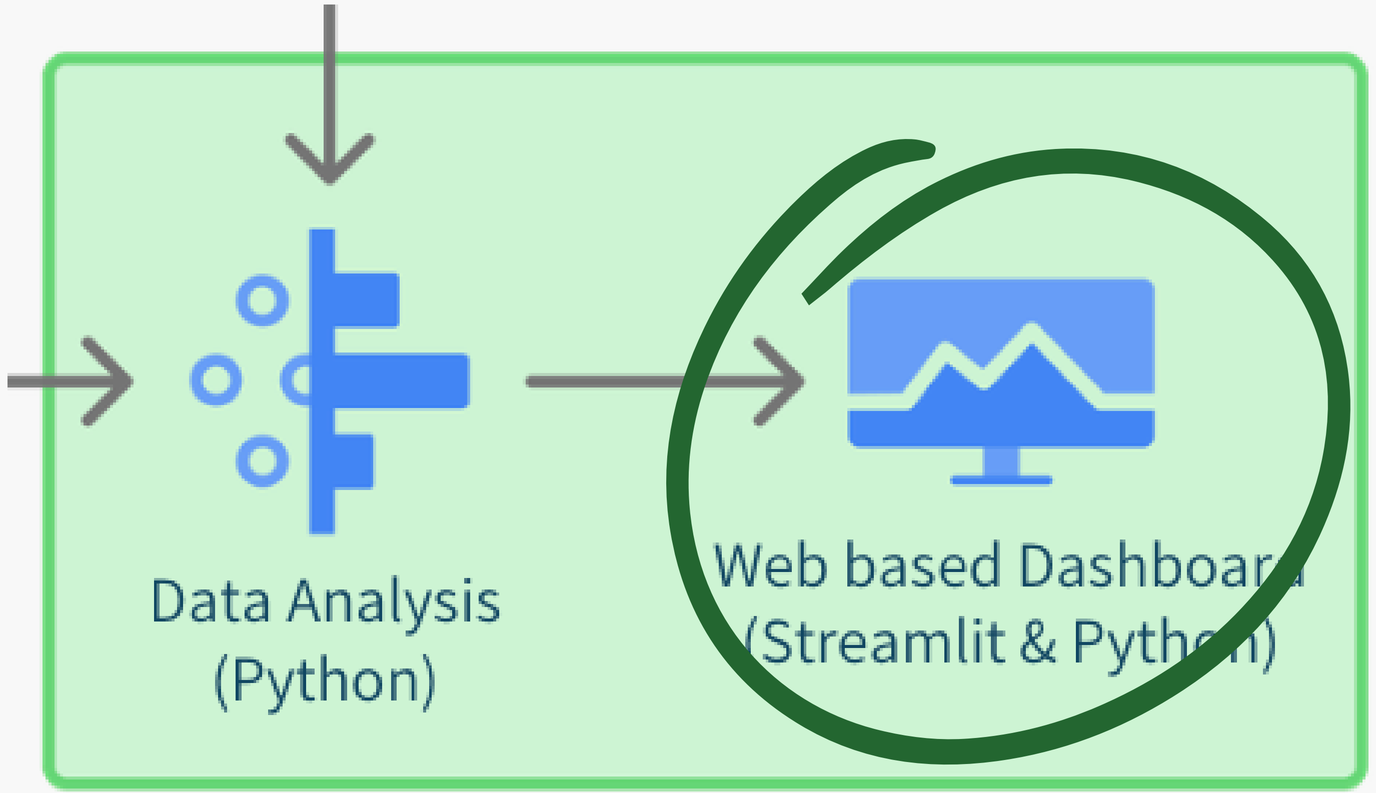
Python

1. Data Integration - Merge national monthly pollutants averages with thermal power generation volume to acquire a unified, 264-month time-series dataset.

2. **Time Series Pattern Analysis**
    a. **Lagged Correlation Analysis:** Determine the optimal time delay where power generation changes most significantly affect PM10 concentration.
    b. **Seasonal Decomposition:** Separate the time series into Trend, Seasonality, and Residuals to control for external factors.

3. **Quantitative Impact Modeling** - Build a Multiple Regression Model to quantify the net effect (coefficient) of the Lag_X power generation on PM10 concentration, controlling for trend and monthly seasonality.

System Architecture - Data Analysis



Streamlit

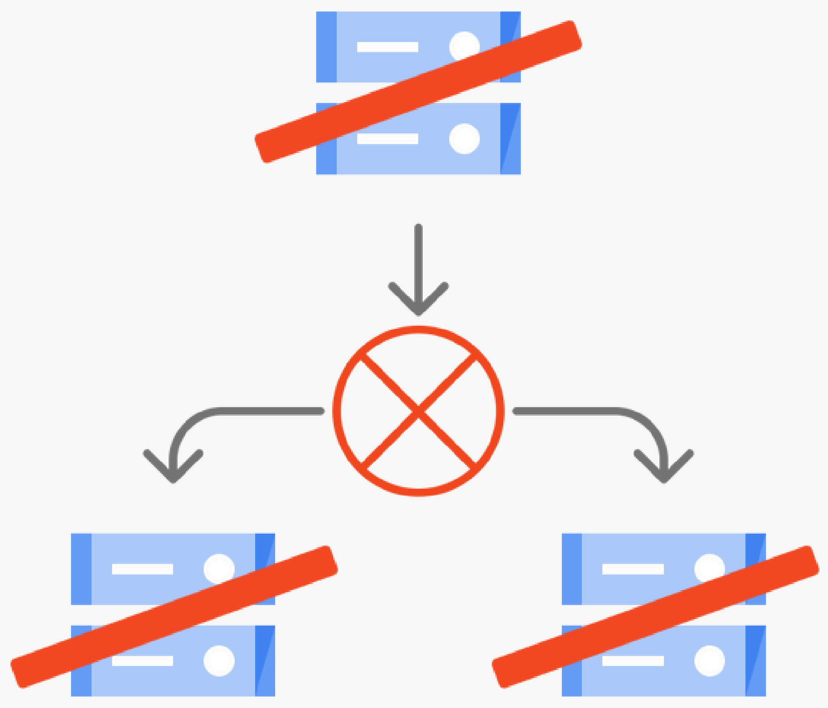
Visualizing Dashboard: Visually see the overall relation between power data, air quality data using correlation analysis, regression analysis, and trend analysis.

Root Cause & Debugging

Cluster Configuration

Objective:

- **3-node Hadoop + Impala + Spark** based Standard structure
- Parquet conversion on **Hive Metastore**, followed by **Impala → Python** integration
- Establish a fully functional **data warehouse architecture** in which HDFS, Hive, and Spark all operate altogether.



Issue encountered: Communication failures between nodes / unassigned DataNodes

- Hostname/hosts **configuration mismatch**
- Unopened Hadoop ports in the **firewall** prevented proper cluster communication.
- As a result, DataNodes were **unable to join** the NameNode, causing the entire pipeline to fail to initialize.

Solution:

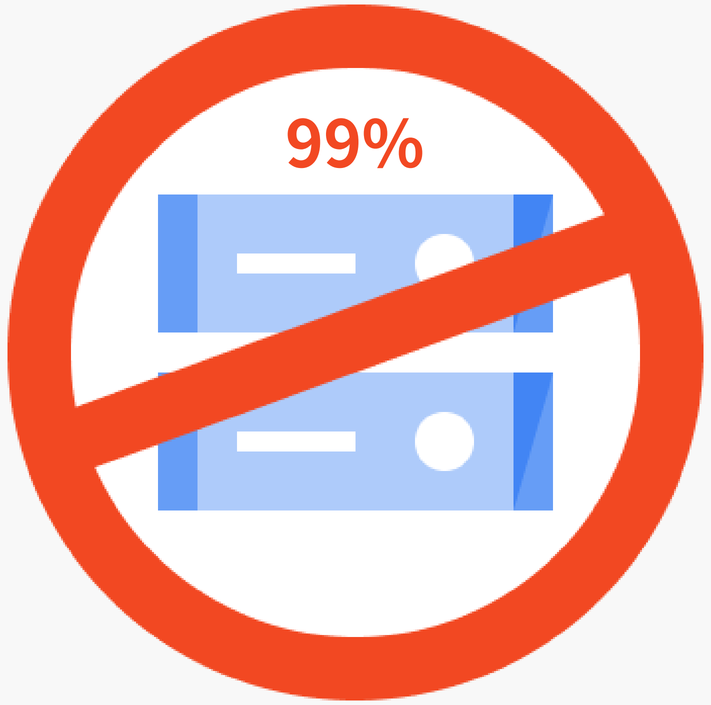
- **Redefined** the hosts, core-site.xml, and hdfs-site.xml **configurations**
- **Disabled network firewall** restrictions to **stabilize internal cluster** communication
- **Verified** the operation of Impala, Hive, and the ResourceManager step by step to **restore system stability**

Root Cause & Debugging

Storage Shortage Issue

Objective:

- Uploaded the entire CSV dataset (approximately 8GB) to the master node's storage, then **distributed** it across HDFS to perform data **preprocessing** and **normalization** using **Spark**



Issue encountered: "The cluster nodes encountered **insufficient storage capacity**"

- Due to the limited default storage allocated in the CDH VM environment, the full dataset **could not be uploaded** to the local storage for processing
- During Spark execution, **out-of-memory** (OOM) errors occurred, causing failures in storing intermediate blocks.

Cause and solution:

- We attempted to **expand** the existing virtual disk (sda), but the GRUB bootloader became corrupted, resulting in a **boot failure**.
- Created a dedicated data storage directory on the master node and **mounted** the new disk (sdb). The directory was then **linked to HDFS** so that it could be recognized and used by the cluster.

Root Cause & Debugging

File Processing Failure Due to Missing CSV Module

Objective:

Preprocessing, normalization, and type casting were performed in **Spark** using its built-in CSV reading and parsing modules.

Issue encountered:

- The provided environment lacked a functional **CSV reader**
- **DataFrame-based method**s failed to process CSV inputs
- The Hadoop environment imposed **restrictions** on native CSV handling

Solution:

- Implemented a **custom parser** for CSV reading and parsing.
- Loaded the CSV file as an RDD and split each row by commas to separate the columns.
- **Manually** constructed the schema.
- Saved the **preprocessed** and **normalized** data in the Parquet format, achieving efficient storage utilization and producing a file structure compatible with Impala.

Root Cause & Debugging

Transition: Hive → Impala

Objective:

- Build a Hive-based data warehouse pipeline: **Spark → Parquet → Hive** external tables.

Issue encountered:

- HiveServer2 **node latency** + **MapReduce overhead** → **severe query slowdown**
- CSV file format → full scans & repeated MR job launches
- Metastore stable, but **network latency + file inefficiency + MR engine** → **slow Beeline responses**

Solution:

- Switch from **Hive to Impala** → immediate performance improvement
- Parquet-native engine → **directly reads** Spark-generated Parquet
- Low-latency DDL → fast metadata loading & table inspection
- Simpler and more stable configuration → fewer errors, more reliable query engine

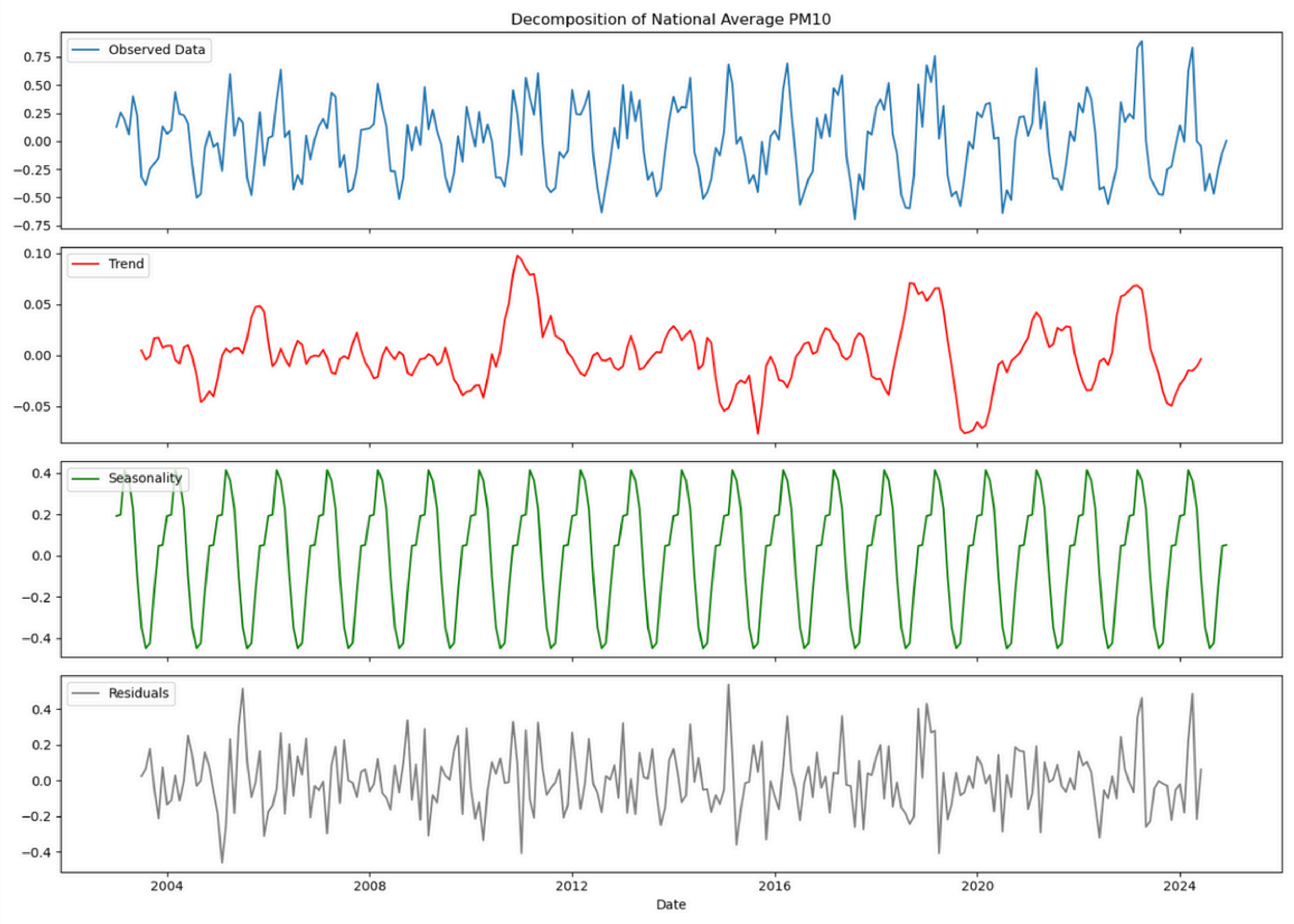
vs.

Results - Lag Correlation

```
=== 2. Start Lagged Correlation Analysis of Power Generation (Power) ===
  - Lag 1 month(s): Correlation = -0.1022, P-value = 0.0982
  - Lag 2 month(s): Correlation = 0.0966, P-value = 0.1187
  - Lag 3 month(s): Correlation = 0.2443, P-value = 0.0001
  - Lag 4 month(s): Correlation = 0.2653, P-value = 0.0000
  - Lag 5 month(s): Correlation = 0.1503, P-value = 0.0155
  - Lag 6 month(s): Correlation = 0.0219, P-value = 0.7265
```

- **Lag 1–2 Months (Short-term): Minimal Impact**
  - The correlation coefficients range from -0.1 to 0.09, and the p-values exceed 0.05.
- **Lag 3–4 Months (Medium-term): Maximized Impact (Key Interval)**
  - Lag 3: Correlation Coefficient 0.244 (P-value 0.0001)
  - Lag 4: Correlation Coefficient 0.265 (P-value 0.0000)
- **Lag 5–6 Months (Long-term): Diminishing Impact**
  - The correlation coefficient drops to 0.15 at Lag 5, and the relationship effectively disappears at Lag 6 with a coefficient of 0.02.

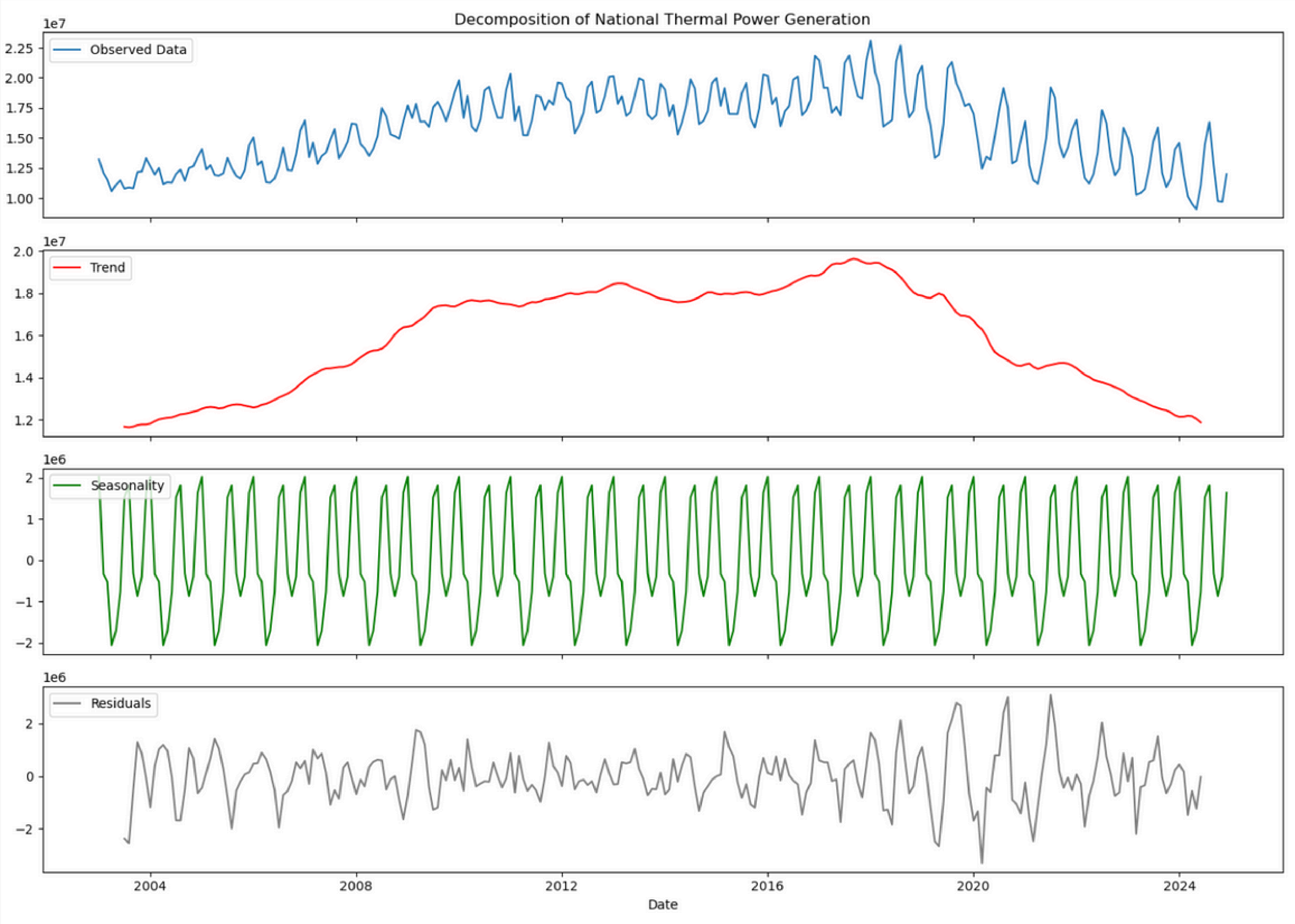
Results - Seasonal Decomposition (1)



Decomposition Results of National Average PM10

- **Trend:**
  - Shows a distinct downward trend, particularly decreasing post-2020.
  - Suggests the long-term effectiveness of government regulations and reduction policies.
- **Seasonal:**
  - Exhibits a clear "Single Peak" pattern.
  - Concentrations spike in Spring (Yellow Dust) and Winter (Heating/Stagnation) while dropping in Summer/Autumn.
- **Residual:**
  - Represents irregular fluctuations excluding trend and seasonality.
  - Spikes likely indicate unexpected anomalies such as massive Yellow Dust events.

Results - Seasonal Decomposition (2)



Decomposition Results of National Thermal Power Generation

- **Trend:**
  - Shows a distinct "Rise then Fall" pattern with a clear inflection point.
  - Steadily increased from 2004 (economic growth), peaked around 2018, and then turned to a decline.
- **Seasonal:**
  - Exhibits a clear "Dual Peak" (M-shaped) pattern, unlike the PM10 data.
  - 1st Peak (Summer): Surge in cooling demand (Jul–Aug).
  - 2nd Peak (Winter): Surge in heating demand (Dec–Jan).
  - Generation drops significantly during low-demand seasons (Spring/Autumn).
- **Residual:**
  - Represents irregular fluctuations excluding trend and seasonality.

Results - Multiple Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      national_avg_PM10   R-squared:                       0.720
Model:                            OLS   Adj. R-squared:                  0.705
Method:                 Least Squares   F-statistic:                     48.68
Date:                Fri, 05 Dec 2025   Prob (F-statistic):           2.01e-60
Time:                        23:50:05   Log-Likelihood:                 79.507
No. Observations:                 260   AIC:                            -131.0
Df Residuals:                     246   BIC:                            -81.16
Df Model:                          13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              0.1529      0.077      1.983      0.048       0.001       0.305
Power_GWh_Lag4  3.314e-09   4.32e-09      0.767      0.444    -5.2e-09    1.18e-08
Trend          -8.303e-05      0.000     -0.535      0.593      -0.000       0.000
Month_2            0.0087      0.057      0.153      0.879      -0.103       0.120
Month_3            0.2221      0.057      3.928      0.000       0.111       0.333
Month_4            0.1656      0.057      2.897      0.004       0.053       0.278
Month_5            0.0349      0.057      0.615      0.539      -0.077       0.146
Month_6           -0.2791      0.056     -4.992      0.000      -0.389      -0.169
Month_7           -0.5436      0.056     -9.722      0.000      -0.654      -0.433
Month_8           -0.6300      0.056    -11.155      0.000      -0.741      -0.519
Month_9           -0.6142      0.056    -10.912      0.000      -0.725      -0.503
Month_10          -0.3662      0.056     -6.543      0.000      -0.476      -0.256
Month_11          -0.1553      0.056     -2.756      0.006      -0.266      -0.044
Month_12          -0.1491      0.057     -2.634      0.009      -0.261      -0.038
==============================================================================
Omnibus:                       11.335   Durbin-Watson:                   1.965
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               11.559
Skew:                           0.483   Prob(JB):                      0.00309
Kurtosis:                       3.368   Cond. No.                     2.05e+08
==============================================================================
```

1. Model Fit

- R-squared (0.720): Explains 72% of variance; indicates high predictive power.
- F-statistic (Prob < 0.05): Confirms the model is statistically valid.

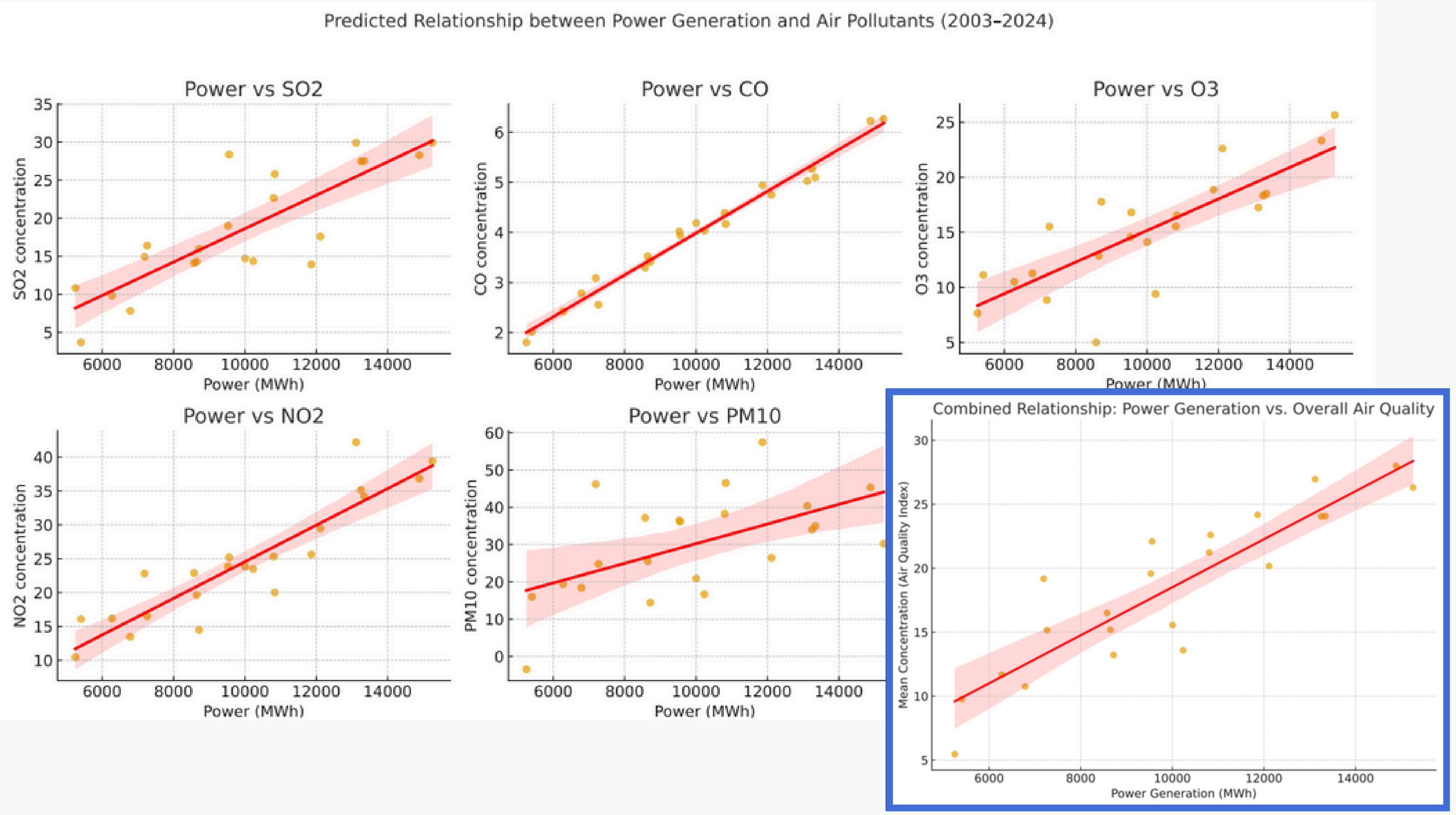
2. Variable Analysis

- Power Generation: Not significant (P-value: 0.444 > 0.05).
- Seasonality: Highly significant (P-value: ~0.000); the dominant factor affecting PM10.

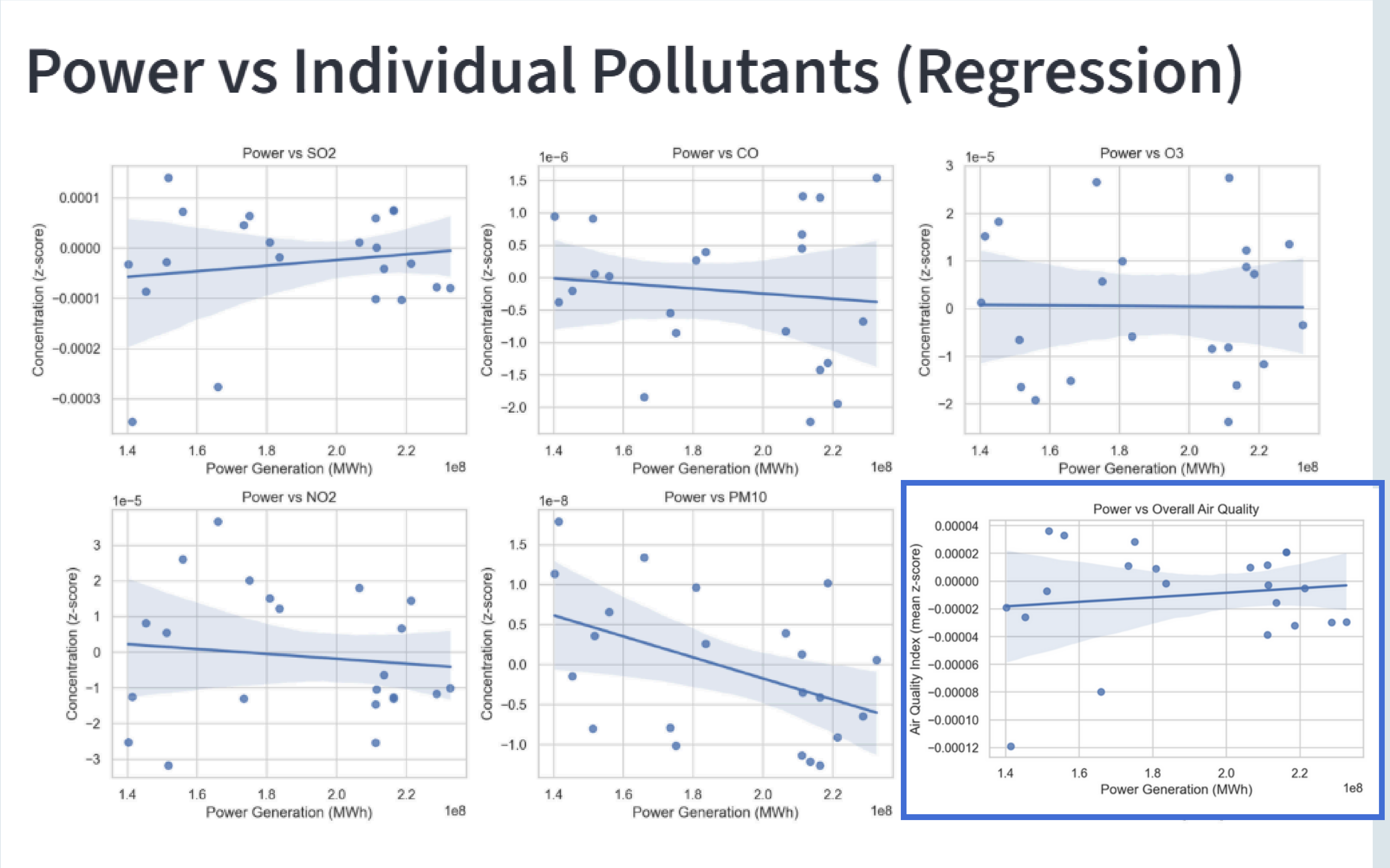
3. Warning

- Multicollinearity: High Condition No. (2.05e8) indicates strong overlap between variables.

Results - Visualization



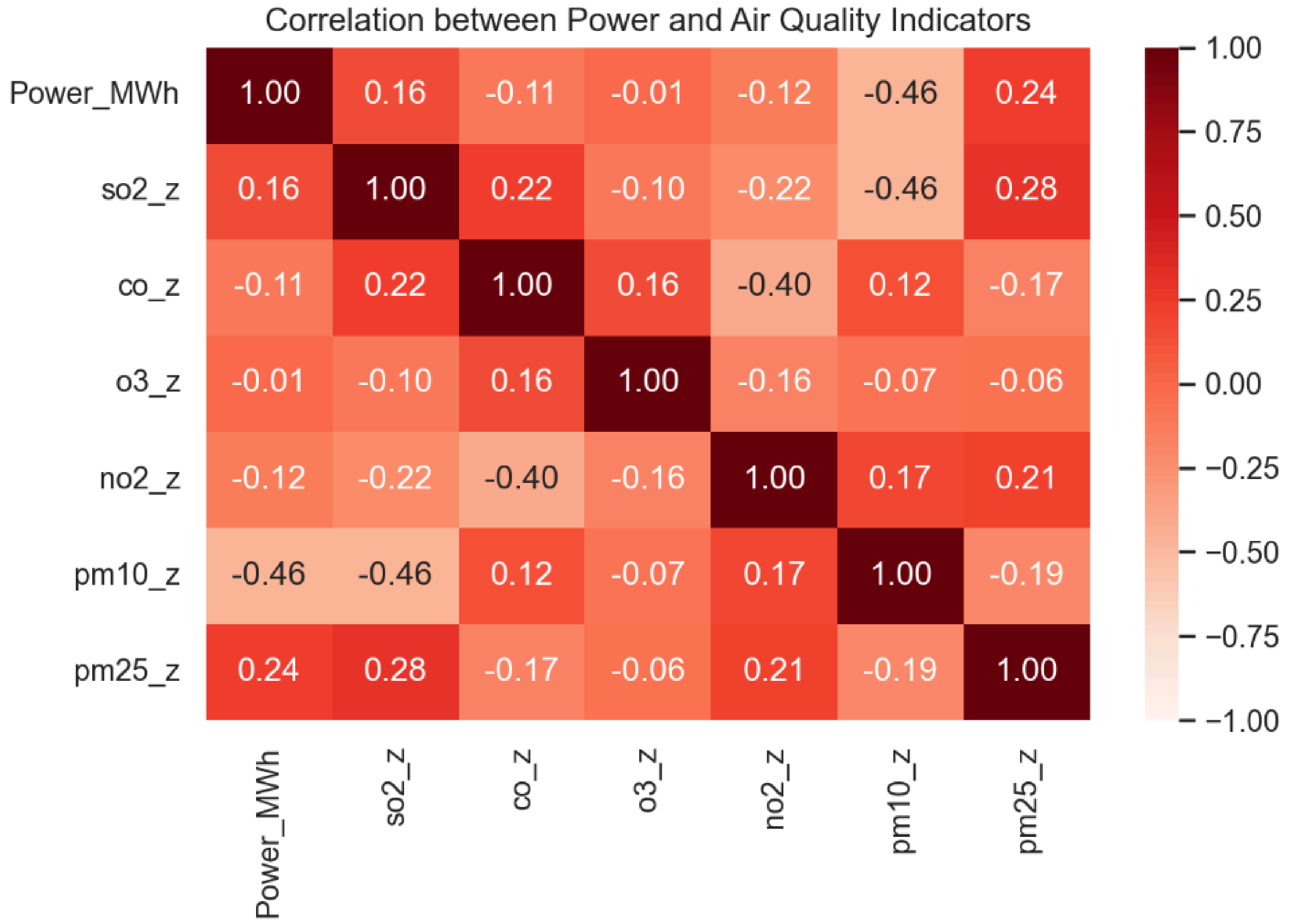
We predicted



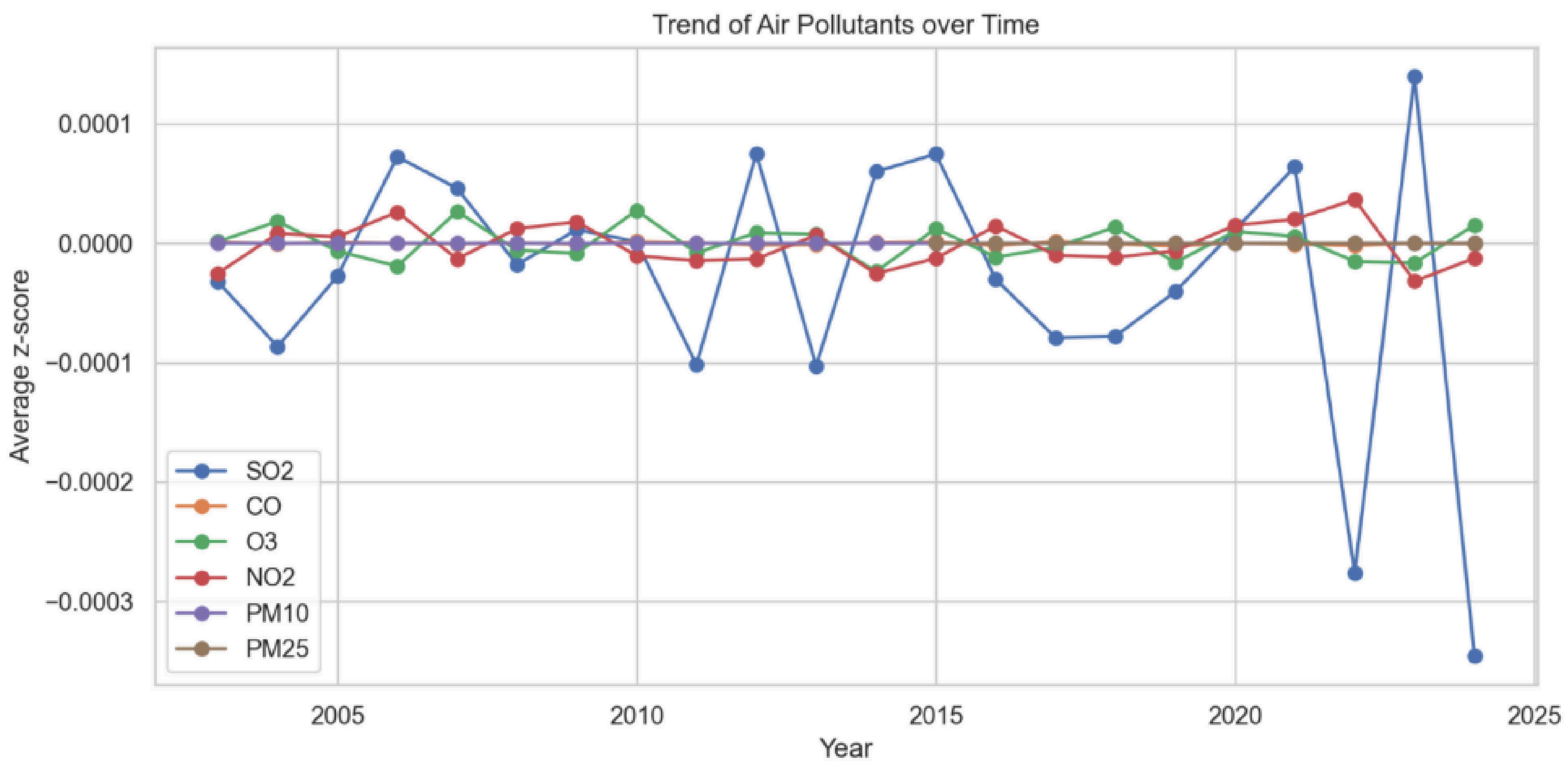
Actual Result

Results - Visualization

Correlation Heatmap



Trend of Air Pollutants over Time



Insights

- Our analysis shows that the observed patterns in air quality indicators are largely shaped by **strong seasonal cycles** and **long-term environmental trends** rather than fluctuations in power generation. Although we evaluated multiple time lags and conducted regression modeling, the statistical evidence consistently indicates that power generation **does not meaningfully explain the variation seen in air quality data.**

- This suggests that Korea’s air quality dynamics are driven by a combination of meteorological conditions, regional pollutant transport, and natural seasonal behaviors. As a result, the relationship between power generation and air quality is inherently **complex and cannot be captured through simple correlations or single-variable analysis.**

✓ Conclusion

- Even though the results did not align with our initial expectations, the project allowed us to go through extensive **trial and error and gain a deep understanding of how an analytical pipeline operates in practice.**

- Through this process, we built a self-updating like analytical pipeline, automating key components such as data collection, loading, and monthly analysis. While the system is not fully intact, we have **established a strong foundation and demonstrated the feasibility of such an approach.**

- Furthermore, we learned that **Mother Nature** is far more **complex** than we can predict, and **air quality cannot be explained by a single variable**. Numerous environmental and meteorological factors influence air quality, meaning that power generation alone cannot serve as a reliable explanatory indicator.

Thank you