Air Quality Data Platform: HORIZON

Samet Erkek, Aysan Pakmanesh, Selin Bozkurt, Melis Ciğer

Istanbul Technical University

APRIL 12 - 2025

Abstract

This project introduces a data-driven platform designed to monitor and analyze air quality using information from both global and local sources. By combining large-scale data processing tools and cloud technologies, the platform transforms raw environmental data into meaningful insights. It allows users to explore air pollution trends, track changes over time, and better understand how air quality is represented in scientific research. The system aims to support more informed decisions in both research and public policy through clear and accessible data visualizations.

1- Introduction & Motivation

Air pollution is a growing concern worldwide, and the availability of open data offers new possibilities to better understand and monitor it. The motivation behind this project was to bring together different types of air quality data from global sources like OpenAQ and OpenAlex to local Turkish monitoring systems and make them easier to process, analyze, and visualize. Handling such large and complex datasets comes with challenges, especially when working with millions of files or real-time data streams. This project aimed to build a cloud-based platform that not only solves these technical problems but also helps researchers and city officials make more informed decisions about air quality.

2- Literature Review

Ambient PM2.5 has gained widespread attention as a major global public health issue, with studies showing that even low levels can pose serious health risks. This review presents updated insights into the mechanisms by which PM2.5 affects human health, contributing to a deeper understanding of its impact. [4]

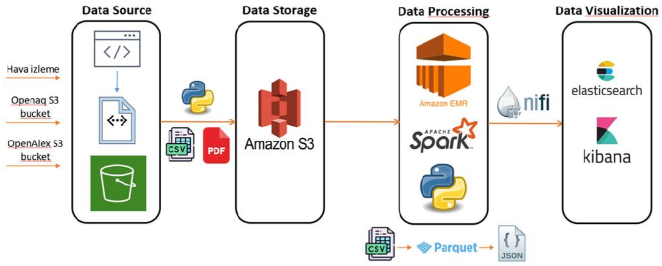
integrating forecasted weather data with existing historical air quality and meteorological records, leveraging machine learning techniques to identify meaningful correlations and develop a highly effective air quality prediction model. The approach focuses on utilizing predictive data features to improve overall forecasting accuracy. As meteorological measurements become increasingly precise and forecast data gains reliability, the potential for valuable insights grows. Effectively combining predictive and historical data can significantly enhance the performance of air quality prediction systems. [1]

The role of geospatial big data in smart cities is becoming increasingly important as urbanization accelerates. Cities now generate vast amounts of spatial data that are essential for improving energy efficiency, controlling pollution, and enhancing overall urban functionality. Geospatial intelligence supports real-time monitoring of environmental changes, traffic patterns, air quality, and the management of smart infrastructure. To meet sustainability goals such as green energy adoption and carbon neutrality, smart cities must leverage advanced data storage, analytics, and visualization technologies. [2]

There is still a lack of a fully integrated real-time system for air quality monitoring and prediction using big data and machine learning. Air quality is influenced by many dynamic factors, requiring robust and efficient models. This paper reviews recent approaches, compares existing systems, highlights key challenges, and suggests practical improvements, noting that current tools often struggle with limited range, accuracy, and efficiency. [3]

3- Technical Approach & Architecture

A cloud-based platform was designed using AWS services (S3, EC2, EMR), Apache Spark for processing, Apache NiFi for data pipelines, Elasticsearch for indexing, and Kibana for visualization. The focus was on handling large-scale data ingestion and processing.



4- Data Source & Storage

OpenAlex (Scholarly PDFs):

A selective download strategy was implemented using EC2 instances to fetch relevant PDF URLs identified through the OpenAlex interface. These files, totaling approximately 868 GB, were processed using a Spark job on Amazon EMR. Since Spark cannot natively read PDFs and listing large numbers of files in S3 can overwhelm the driver, this step required careful memory management. The extracted text content was structured into DataFrames and saved as optimized Parquet files (10 GB total) using .coalesce() to reduce file fragmentation.

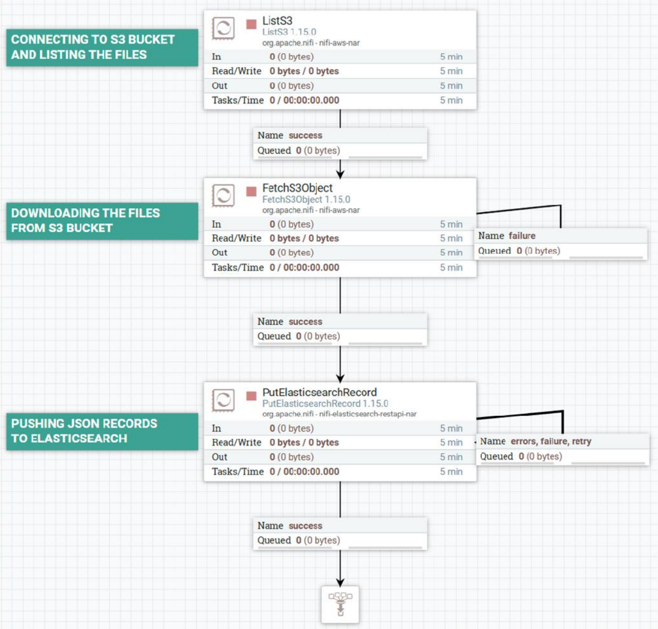
OpenAQ (Global Air Quality Data):

Python scripts were developed to fetch year-by-year data from the OpenAQ S3 bucket and copy it into the project's own S3 bucket. This process ran in parallel across small EC2 instances to speed up ingestion. The raw dataset contained 28 million small compressed CSV.GZ files (19.3 GB total), which were unsuitable for direct large-scale processing due to the "small files problem" common in Spark.

TR Hava İzleme (Turkish Monitoring System):

Historical air quality data (originally in XLSX format, converted to CSV) was ingested through a NiFi pipeline using processors like ListS3, FetchS3Object, and ConvertRecord.

Real-time data was collected hourly via a local Python script due to the lack of an API and possible IP-based restrictions. The data was sent directly to Elasticsearch.


CONNECTING TO S3 BUCKET AND LISTING THE FILES
DOWNLOADING THE FILES FROM S3 BUCKET
PUSHING JSON RECORDS TO ELASTICSEARCH

ListS3
ListS3 1.15.0
org.apache.nifi - nifi-aws-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

FetchS3Object
FetchS3Object 1.15.0
org.apache.nifi - nifi-aws-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name failure
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

PutElasticsearchRecord
PutElasticsearchRecord 1.15.0
org.apache.nifi - nifi-elasticsearch-restapi-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name errors, failure, retry
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

5- Data Processing

OpenAQ

In the processing part of the project it has been focused on transforming raw data into an analytics-ready format while improving data quality. A Spark job was executed on a 10-core EMR cluster to convert the raw OpenAQ CSV files into Parquet format. To avoid overwhelming the Spark driver, a separate Python script was used to pre-list all file paths into a CSV, which Spark then read efficiently.

This process allowed the entire dataset to be converted in approximately 8 hours. After conversion, the Parquet outputs were merged into a smaller set of about 22 files (each around 300 MB) to optimize read performance in subsequent queries.

Following this, Spark was used for data cleaning and enrichment. Invalid measurements (e.g., negative values, missing values) were removed and new columns such as year, month, and day were extracted. Since the 'location' field in the OpenAQ dataset was inconsistent which didn’t have any specific dinformation which it was unclear, the data was enriched by joining it with the GeoNames dataset using latitude and longitude values. This integration added standardized country names, city names, and population data, resulting in a cleaner, reliable and more understandable dataset.

OpenALEX

To understand how air quality topics appear in academic research, a PySpark job was run on the OpenAlex data that was stored in s3 bucket. The text was first cleaned by making everything lowercase and removing punctuation. Then it was split into individual words, and common or meaningless terms like numbers and typical stop words were filtered out. This helped focus on the actual keywords researchers tend to use in their publications.

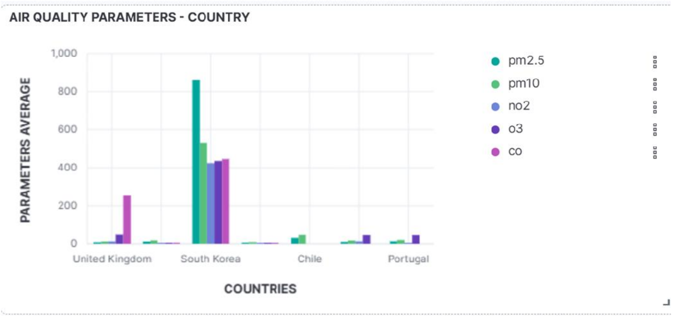
After cleaning, the most frequently used words were counted and the top 100 were saved as a single JSON file in S3, ready for visualization in Kibana. Similar jobs also looked at how often specific years or terms like pm25, co2, and ozone appeared in the texts. Running these jobs on an EMR cluster made it possible to process large volumes of data quickly and get a clearer picture of how air quality is reflected in the scientific literature.

Considering all the processes that has been passed through. The Bronze layer represented the raw OpenAQ data (600 GB), while the Silver layer consisted of cleaned and partitioned data (300 GB). The final Gold layer included filtered 2024 data converted to JSON format (8 GB) for visualization purposes. An Apache NiFi pipeline was established to list , download these JSON files in S3 and push them into Elasticsearch, enabling both real-time and historical visualizations through Kibana dashboards.

6- Data Visualization

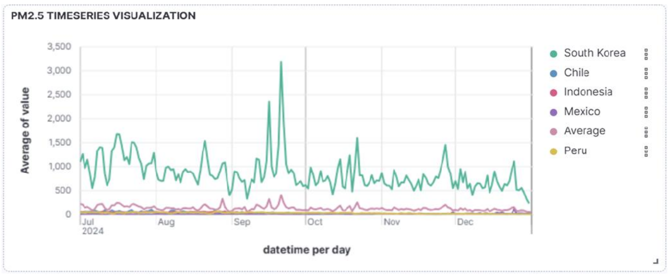
The final data was visualized using Kibana. Kibana enabled the creation of interactive dashboards to monitor air quality metrics such as PM2.5, NO_2, and O_3 across different regions and time periods. These visualizations provided valuable insights for identifying pollutional problems and tracking trends in air quality.

AIR QUALITY PARAMETERS BY COUNTRY



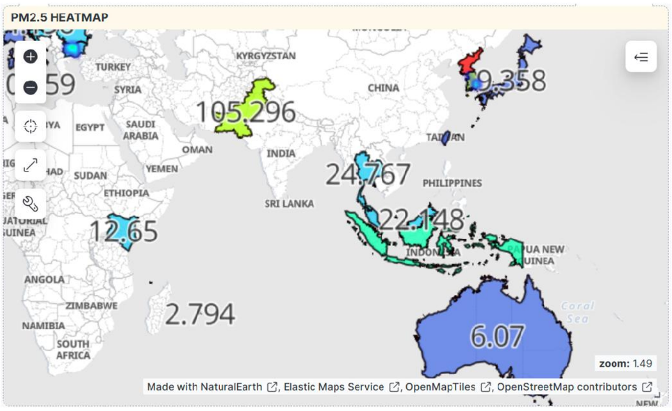
This bar chart provides the air quality parameter levels across different countries. Each colorcoded bar represents a specific pollutant. South Korea has the highest levels across all air quality parameters, especially PM2.5 and PM10, indicating severe pollution. In contrast, Portugal and Chile show significantly lower values, reflecting better air quality conditions.

PM2.5 TIMESERIES VISUAL BY COUNTRY



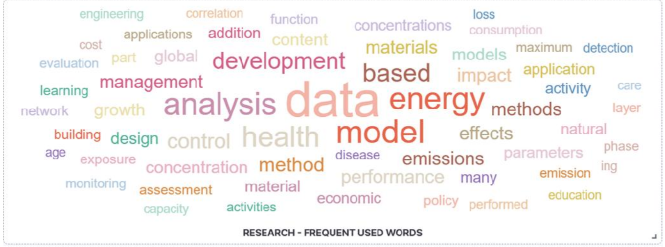
This time series chart shows daily average PM2.5 levels from July to December 2024 across six countries. South Korea consistently records significantly higher PM2.5 values compared to others, with several sharp edges in pollution levels. In contrast, countries like Chile, Mexico, and Peru maintain relatively stable and low PM2.5 levels comparing to both average and South Korea values.

PM2.5 HEATMAP



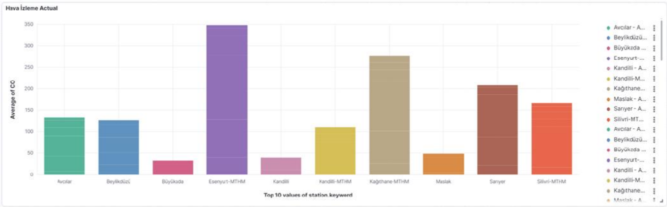
This PM2.5 heatmap displays the average pollution levels across various countries. PM2.5 levels are high in both North and South Korea due to industrial activities, coal-based energy use, and high traffic emissions. In South Korea, pollution is further worsened by transboundary dust, fine particles carried from China and traffic. Additionally the lack of environmental regulations also contribute significantly to elevated PM2.5 levels.

WORD CLOUD VISUAL OF RESEARCHS



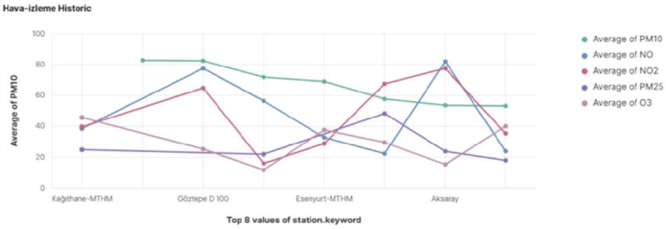
This word cloud suggests that air quality research commonly explores the health impacts of pollution, with keywords like health, exposure, and disease standing out. Terms such as energy, emissions, and consumption point to the influence of industrial activities and fuel use on air quality. Meanwhile, the presence of management, control, and policy reflects the importance of regulation and strategies aimed at monitoring and improving environmental conditions. For future studies; features can be extracted from this word cloud visualization or subjects that will correlate with air quality can be defined for new researches.

STREAMING PARAMETER VISUAL OF ISTANBUL DISTRICTS



This visualization shows the real-time average levels of air quality parameters across different districts in Istanbul. Stations like Esenyurt and Kağıthane report significantly higher pollution values, indicating potential hotspots that may require immediate attention. Such insights can help local authorities prioritize actions and allocate resources more effectively for air quality improvement.

HISTORIC PARAMETER VISUAL OF ISTANBUL DISTRICTS



By combining real-time and historical air quality data across Istanbul districts, it's clear that areas like Esenyurt and Kağıthane consistently show higher pollutant levels such as PM10 and NO₂. The historical trends reveal fluctuating values in stations like Aksaray and Göztepe, suggesting episodic pollution spikes, while others like Şirinevler and Sultangazi show more stable but still concerning averages. These insights can guide municipal authorities in identifying both persistent problem zones and temporary pollution events, helping to shape targeted air quality interventions.

7- Future Considerations

Planned future work includes automating ingestion pipelines, integrating machine learning algorithms, expanding data coverage for the Turkish system, further optimizing the OpenAQ processing, potentially migrating local components to the cloud.

8- Conclusion

In essence, this platform offers a comprehensive and scalable solution for monitoring and analyzing air quality by integrating diverse data sources and technologies. Through the combined use of global datasets like OpenAQ and OpenAlex, local monitoring systems, and a cloud-based architecture powered by AWS, the system ensures reliable data ingestion, transformation, and enrichment. Apache Spark and EMR enable efficient large-scale processing, while NiFi pipelines automate data flow toward Elasticsearch for indexing. Finally, with Kibana dashboards, users can interactively explore real-time and historical trends, gaining actionable insights into pollution patterns at both global and local scales. This project bridges the gap between raw environmental data and meaningful interpretation, empowering researchers, policymakers, and city officials with the tools needed for informed decision-making.

References:

[1] Y. Zhang et al., "A Predictive Data Feature Exploration-Based Air Quality Prediction Approach," in IEEE Access, vol. 7, pp. 30732-30743, 2019, doi: 10.1109/ACCESS.2019.2897754.

[2] Mete, M. O.: GEOSPATIAL BIG DATA ANALYTICS FOR SUSTAINABLE SMART CITIES, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-4/W7-2023, 141–146, https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-141-2023, 2023.

[3] Gangwar, A., Singh, S., Mishra, R. et al. The State-of-the-Art in Air Pollution Monitoring and Forecasting Systems Using IoT, Big Data, and Machine Learning. Wireless Pers Commun 130, 1699–1729 (2023). https://doi.org/10.1007/s11277-023-10351-1

[4] Shaolong Feng, Dan Gao, Fen Liao, Furong Zhou, Xinming Wang, The health effects of ambient PM2.5 and potential mechanisms,Ecotoxicology and Environmental Safety,Volume 128,2016,Pages 67-74, ISSN 0147-6513, https://doi.org/10.1016/j.ecoenv.2016.01.030