A Big Data approach to Air Quality Monitoring in Cotabato City

1st Julienne Kate N. Kintanar
Kidapawan City, Philippines
Mindanao State University - Iligan Institute of Technology
School of Interdisciplinary Studies
juliennekate.kintanar@g.msuiit.edu.ph

2nd Abdul Mojeer M. Akmad
Cotabato City, Philippines
Mindanao State University - Iligan Institute of Technology
School of Interdisciplinary Studies
abdulmojeer.akmad@g.msuiit.edu.ph

***Abstract*—Air pollution has become a global concern especially in urban areas causing public health problems such as respiratory and cardiovascular illnesses. In the Philippines, the government has enacted the Clean Air Act to establish standards and monitoring systems. The BARMM government has initiated the establishment of air quality monitoring stations in Cotabato City to preserve and protect the air quality in the region. This paper proposes the integration of Big Data framework and tools like Resilient Distributed Dataset, DataFrames, PySpark, Hadoop Distributed Filing System and Parquet to process, analyze and visualize the data. The objective of this paper is to collect, store and process the data to create visualization that will enable real-time monitoring that will support timely health advisories and evidence-based policy formulation. It will also become the foundation framework for further expansion of other monitoring stations in the region. The expected outcome of this proposal will foster better public awareness and contribute to a more healthy and sustainable environment.**

***Index Terms*—Big Data Approach, Air Quality, Spark, Data Storage**

I. Introduction

Due to rapid urbanization and industrialization, environmental pollution issues such as air pollution have become more and more severe. Environmental pollution is skyrocketing and many diseases, especially respiratory, cardiovascular, and tumors are affecting the population [1]. Several studies have strongly linked air pollutants to asthma and lung disease]. A 2016 report of the World Health Organization has stated that 92% of the global population live in regions where the air quality level exceeds the limits which have caused approximately 3 million deaths annually due to air pollution related diseases [2]. In the Philippines, the air quality problem cost 66,320 lives and an estimated PHP 2.32 trillion in economic costs [3].

The Philippines have enacted the Republic Act 8749 otherwise known as Philippine Clean Air Act of 1999 to protect and preserve the air quality by establishing emission standards, penalizing violators and putting up Air Quality Management and Monitoring Systems [4]. The air quality monitoring system (AQMS) is established to collect and gather data and information on the current ambient air quality and pollutants in order to provide timely and accurate for policymakers in making policies and the public at risk in protecting their health. The major pollutants include ozone, particulate matter, sulfur dioxide, carbon monoxide, and nitrogen oxides among others [5].

The BARMM region through the Ministry of Environment Natural Resources and Energy has started the establishment of AQMS starting in Cotabato City where the seat of regional government is located. The BARMM government is also planning to establish AQMS in every major city and town of the region. The AQMS in the City was inaugurated in October 2023 and started the collection of hourly air quality data for public dissemination [6].

The continuous monitoring and collection of data leads to high volume data with spatial nature. Analyzing and predicting air pollution using these datasets requires data processing technologies with high processing and storage capacity [7]. Traditional data analytics cannot keep pace with the requirements of high volume and variable data generated by AQMS. This paper will discuss the application of Big Data technology specifically the use of Spark tools such as Resilient Distributed Dataset, DataFrames, PySpark, Hadoop Distributed File System and Parquet format to process, store and analyze data from the Cotabato City AQMS.

II. Project Details

A. Problem Statement

Cotabato City, like other cities, experiences the problem of increasing air pollution, which is a major concern in the field of health as well as environment. Traditional analysis techniques are unable to efficiently store and process the huge volumes of data produced by the Air Quality Monitoring System (AQMS) at an hourly rate. This research seeks to overcome this limitation by developing Big Data solutions particularly Apache Spark to process air quality data and extract insights from the variables available.

B. Objectives

This study aims to explore the application of Big Data technology specifically the Spark software in the analysis of air quality data of Cotabato City. Furthermore, the paper specifically aims to:

1. To collect, store and integrate the hourly air quality data of Cotabato City AQMS using Spark
2. To apply data processing and analysis techniques to manage big volumes of data.
3. To create visualization to effectively communicate the trends and patterns to the public, policymakers and stakeholders.
4. To evaluate the feasibility of expanding the scope for future AQMS in other major towns of BARMM.

III. Big Data Concepts and Tools

A. Resilient Distributed Dataset

The RDDs' ability to perform parallel processing is one of the necessary features in handling large datasets. In the context of air quality monitoring, where data is gathered at consistent intervals (hourly rate), RDDs can distribute the tasks across multiple nodes within a cluster. The capability of RDD can significantly speed up the computations required to examine variations in pollutant levels in connection with weather conditions [8]. RDDs can essentially handle unstructured or semi-structured data, including those datasets with various formats and types of measurements. Moreover the inferent fault tolerance, one of the strengths of RDD that allows Spark to recompute lost data partitions in case of node failures, ensures that air quality data remains accessible and reliable for the ongoing analysis, which is particularly important in real-time monitoring.

B. DataFrame

Spark DataFrame is a structured collection of data organized into rows and columns that resembles traditional databases, which makes it easy to work on cleaning, transforming, aggregating and analyzing our air quality dataset for patterns/trends/anomalies/missing values. In addition, data can be visualized with Python libraries such as Matplotlib and seaborn that are both attractive and useful for informative visualization. Data can be provided to Spark DataFrames in many forms (including CSV), and they are capable of dealing with data sizes ranging from megabytes up into the petabyte range, as well interfacing with a broad variety of Big Data tools [9].

C. PySpark

PySpark performs complex data processing operations with the help of Dataframe API and Resilient Distributed Datasets (RDDs). In processing available data on air quality, PySpark can execute transformations like filtering, aggregation and joining of the different datasets effectively because of the large volume of data. For example, it is possible to determine average pollutant concentration levels over periods of time or the changes of indices of air pollution using built in functions [10]. PySpark includes a wide range of statistical functions that can be implemented to achieve EDA. The describe() function on DataFrames can be used to estimate mean, median, standard deviation and percentile. Further, the visualizations can be done by transforming PySpark DataFrames into Pandas DataFrames for further processing using Matplotlib or Seaborn [11]. For example, to represent changes in pollutant concentrations, which can be done with line charts, or to compare density of different AQ indices with histograms or box-plots. For additional analytical computations, PySpark is compatible with other machine learning libraries including MLlib. It is possible to use regression models to estimate indices of air quality, relying on data from the past. For instance, linear regression can be applied in the prediction of the general AQI depending on the various pollutants [12].

D. Hadoop Distributed File System

HDFS can be used with efficiency in managing and storing air quality dataset which encompasses various environmental factors like wind speed, humidity and air quality indices. HDFS is designed in a way to deal with mass data across a distributed system. With an increase in the size of the air quality dataset, which may include more variables, or a higher frequency of measurements overtime, HDFS can expand by simply adding more nodes to the cluster. This implies that as the data grows, we do not have to make major adjustments to the data architecture [13]. HDFS is capable of offering high throughput on application data and that makes it suitable for big data analytical tasks. For instance, while working with trends in air quality or when performing batch analysis of historic data HDFS provides high performance for reading and writing due to the distributed architecture of the data storage. This capability greatly helps in cutting down the time needed for the analysis using data processing methods for handling large amounts of air quality data [12].

E. Parquet

Parquet is a columnar storage file format that can effectively provide the best storage as well as query optimization. The column-based storage design decreases disk usage while at the same time allowing easy access to particular columns when performing analysis. Concerning this research, using Parquet format to store air quality data in HDFS will improve the general performance of data processing. Due to its integration with Spark's DataFrame API, Parquet enhances structured data, and enhances the ability to query and analyze large datasets. This capability is vital for producing timely analysis of the trends and patterns of air quality to be used in the formulation of the public health measures [–14].

IV. Data Collection and Storage Plan

The dataset for this proposal will be sourced from the Air Quality Monitoring Station established in Cotabato City which is managed by the Ministry of Environment, Natural Resources and Energy - BARMM in partnership with BP Integrated Technologies Incorporated. The dataset includes meteorological data such as temperature, humidity, wind speed, and wind direction. It also includes pollutant concentrations such as ozone, particulate matter, sulfur dioxide, carbon monoxide, nitrogen dioxide and nitrogen oxide. The collected data will be ingested into the Hadoop Distributed File System (HDFS), which provides a scalable and fault-tolerant storage solution. The raw data will initially be stored in CSV format for ease of ingestion. However, for optimized storage and query performance, processed datasets will be converted into Parquet format to allow for efficient compression and faster access to specific columns during analysis.

V. Data Processing Approach

A. Phase 1: Data Cleaning and Preprocessing

After the raw data is stored in HDFS, initial processing will be performed loading data into Resilient Distributed Datasets (RDDs). RDD is a low level transformation that makes it best suitable for unstructured or semi-Structured data. At this stage, we operate by filtering invalid records or applying transformations to treat inconsistent data. This step cleans the data and makes it ready for further structured Analysis.

When initial cleaning is done, we will convert our processed RDDs into PySpark DataFrames. This transition gives a way to make use of the higher level abstractions offered by DataFrames which are structured collections of data including named columns. DataFrame will give advantages in terms of structured data query and data manipulation process along with data aggregation.

B. Phase 2: Data Analysis and Exploration

With the dataset in DataFrame format, advanced analytical operations that will meet the objectives of the project will now be performed by using Pyspark. This comprises using operations in-built in the various analytical tools to perform statistical analysis like computing average concentration of the pollutants over the various intervals of time and exploratory data analysis (EDA) to gain insights on the air quality. The ability to visualize these trends through statistical functions enhances understanding and communication of results to stakeholders.

Throughout the analysis process, both initial results and final datasets will be stored in Parquet format within HDFS. Parquet's columnar storage optimizes both storage space and query performance, allowing for faster access to specific columns during analysis. This capability is efficient in handling the AQMS data as more monitoring stations are added across the region.

C. Phase 3: Data Visualization

Using PySpark DataFrames, the processed air quality data will be transformed and displayed through a variety of additional libraries, specifically Matplotlib or Seaborn. These libraries are essential for creating visual representations of the data, which will be incorporated into dashboards or reports. The main goal of these visual representations is to provide a clear and detailed view of changes in air quality over time. Different types of graphical formats, including line charts, histograms, and other charts that are best suited to present data. Through this, stakeholders in Cotabato city will be able to see the trend or any patterns in the data. In that way, they will be equipped to take proactive steps to address air pollution issues effectively.

D. Phase 4: Interpretation

Interpreting the visualized data to acquire potential actionable insights that will inform the local authorities on the trend and status of air quality in the City will be done in this phase. Policy adjustments and practical steps may be taken on the concerned group of people when the results interpret high risk of air pollution on public health. In pursuit of the research objectives, evaluation will be conducted assessing the feasibility of expanding the AQMS framework to other major towns within BARMM. This assessment will consider scalability aspects of both the data processing infrastructure and its applicability to similar contexts in different locations.

VI. Expected Outcomes and Impact

The implementation of the proposal is projected to demonstrate several outcomes and potential impacts to the policymakers, industries, and citizens of Cotabato City. The following are the expected outcomes and potential impacts:

- Enhanced real-time air quality monitoring: the current practice uses traditional reporting of sensor results that results in frequent maintenance and not updated advisories. The integration of the big data framework will allow continuous monitoring of air quality in the City. The timely information will enable the real time advisories and alerts on pollution level informing local authorities and the public to take precaution as they happen.
- Comprehensive Data Visualization: the current reporting practice is limited in the hourly posting of air quality parameters only. The use of big data will enable the data to be visualized in the form of a dashboard containing charts and plots to illustrate the concentration changes over time.
- Policy Support: the data from the air quality monitoring station can be used for supporting the creation of AirShed Governing Boards and the declaration of attainment and nonattainment areas as prescribed by the Philippine Clean Air Act. The board or the local authorities can use the data to craft laws that are evidence-based to help in public health and control the emissions of industries or transportation vehicles.
- Scalable to the whole region: the proposal will first cater with the data from Cotabato City but the framework shall be established to handle more data that will come in the future monitoring stations. The scalability will enable the regional approach in the policy-making and management of air quality.
- Improvement in Public Health and Productivity: early warnings on the pollution levels will enable the public to take actions to avoid going out or use protective masks. This is very important to groups with health sensitivities

or vulnerabilities such as people with respiratory and heart diseases. Healthcare problems will lead to lower productivity of the workforce, reducing these scenarios will boost productivity and ensure the health and safety of the public.

References

[1] J. Wang, J. Zhang, X. Yuan, Y. Tang, H. Hao, Y. Zuo, and H. Chen, "Air quality data analysis and forecasting platform based on big data," in *2019 Chinese Automation Congress (CAC)*. IEEE, November 2019, pp. 2042–2046.

[2] D. H. Shih, P. Y. Shih, and T. W. Wu, "An infrastructure of multi-pollutant air quality deterioration early warning system in spark platform," in *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, April 2018, pp. 648–652.

[3] L. Myllyvirta, H. Thieriot, and I. Suarez, "Estimating the health & economic cost of air pollution in the philippines," February 6 2023, retrieved from https://energyandcleanair.org/publication/cost-of-air-pollution-in-the-philippines/.

[4] "Philippine clean air act of 1999, republic act no. 8749," June 23 1999, faolex Database. Entry into force 15 days after publication. Available from https://www.fao.org/faolex/results/details/en/c/LEX-FAOC045271/.

[5] Department of Environment and Natural Resources (DENR), "Air quality monitoring is a top priority," March 15 2020, retrieved from https://denr.gov.ph/news-events/denr-air-quality-monitoring-is-a-top-priority/.

[6] Environmental Management Services, "1st air quality monitoring station in barmm, new menre executive building inaugurated," October 5 2023, retrieved from https://menre.bangsamoro.gov.ph/1st-air-quality-monitoring-station-in-barmm-new-menre-executive-building-inaugurated/.

[7] J. Gonzalez *et al.*, *Apache Spark: The Definitive Guide*. O'Reilly Media, 2016.

[8] M. Asgari, M. Farnaghi, and Z. Ghaemi, "Predictive mapping of urban air pollution using apache spark on a hadoop cluster," in *Proceedings of the 2017 International Conference on Cloud and Big Data Computing*, September 2017, pp. 89–93.

[9] Microsoft, "Tutorial: Load and transform data using apache spark dataframes," August 29 2024, available at https://learn.microsoft.com/en-us/azure/databricks/getting-started/dataframes.

[10] Y. Li, Y. Wang, and X. Chen, "Big data analytics for air quality monitoring: A review," *Environmental Science & Technology*, vol. 54, no. 12, pp. 7467–7480, 2020.

[11] A. Kharwal, "Air quality index analysis using python," 2023, retrieved from https://thecleverprogrammer.com/2023/09/18/air-quality-index-analysis-using-python/.

[12] GeeksforGeeks, "Predicting air quality index using python," 2024, retrieved from https://www.geeksforgeeks.org/predicting-air-quality-index-using-python/.

[13] Integrate.io, "The ultimate guide to hdfs for big data processing," 2023, retrieved from https://www.integrate.io/blog/guide-to-hdfs-for-big-data-processing/.