Data Engineering Pipeline to Analyse Jakarta’s Air Quality during COVID-19-Caused Lockdown Periods

Reza Rahutomo1,3, Bens Pardamean2,3
1Information System Department, School of Information Systems Bina Nusantara University Jakarta, Indonesia 11480
2Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia 11480
3Bioinformatics & Data Science Research Center Bina Nusantara University Jakarta, Indonesia 11480
E-mail: reza.rahutomo@binus.edu

Abstract. Jakarta lifted up lockdown after passing more than 50 days of large-scale social activity restriction and initiated phase opening to new normal. To analyse Jakarta’s air quality after passing lockdown, a pipeline of data engineering is needed. By acquiring time series data from openaq.com, a time-series database system is developed with Python programming language and its fundamental libraries namely Pandas, NumPy, SQLite. After PM 2.5 data are pre-processed into average per-hour and grouped by applicable periods (pre-lockdown, lockdown, and phase opening), a pattern of PM 2.5 in South Jakarta is revealed by using data visualization library Matplotlib. The apex of PM 2.5 occurs earlier during lockdown (04:00) and phase opening (02:00) rather than when it was normal or pre-lockdown (08:00) even though the nadir of PM2.5 still occurs at the same time (16:00 – 17:00).

Keywords: Data Engineering, Data Pipeline, Jakarta, Air Quality, PM 2.5

1. Introduction
After more than 50-day lockdown or locally known as the term PSBB (Pembatasan Sosial Berskala Besar), the Provincial Government of Jakarta initiated a relaxation as a phase opening to new normal starting from June 5th, 2020 [1]. The strategy was taken to increase the mobility of residents which will strengthen the economy even though the spread of Corona Virus Disease 19 (COVID-19) is not over yet [2]. This strategy was drafted after several assessments were applied to various indicators in Jakarta, namely: Epidemiology, Public Health Level, and Health Facilities. Since Jakarta’s readiness score is more than 70, the city has a big opportunity to establish a transition and to recover the people’s economy[3].

During phase opening, the local government allowed worship places, shopping centres, and offices back to operation with health protocol applied [4–6]. But not for schools and universities that still rely on online learning as the Ministry of Education ordered educational institutions to remain closed to avoid the greater risk of the virus spreading within students. In this situation, online learning has advantages such as fostering a positive feeling in IT infrastructure and improving learning desires for further sessions [7]. Even so, it requires content updates and computer skills to produce a satisfying experience [8, 9].
Rahutomo, Purwandari, Sigalingging, and Pardamean studied Jakarta’s air quality and discovered that lockdown affected the concentration of particulate matter 2.5 (PM 2.5) according to Air Quality Index (AQI) classification [10]. Since lockdown was established, the concentration of PM 2.5 had decreased to the lower range of unhealthy level compared to pre-lockdown where it remained stable at the upper range of unhealthy level.

PM 2.5 has a unique characteristic that generally higher in the time when the temperature drops like midnight or winter than when the temperature rises like in a rush-hour or summer [11, 12]. Yang, Yuan, Li, Shen, and Zhang reviewed that the rise and fall of PM 2.5 are not always related to emissions but also environmental and geographical factors namely topography, vegetation, and climate. It positively correlated with air pressure, but negatively correlated with wind speed, temperature, humidity, and precipitation[13].

This study emphasized in developing a data engineering pipeline to facilitate PM 2.5 analysis during lockdown phases in Jakarta. Data engineering is foundation to work with big data, and data pipeline is an element of data engineering that allows data movement from one or multiple sources to local or cloud server for analysis and visualisation [14, 15]. If time series data of PM 2.5 managed with the right protocol, in the future it will be highly usable to develop various applications namely time-series regression [16, 17], classification system [18], recommendation system [19], recognition system [20], even monitoring and evaluation system [21]. South Jakarta is the area where the study is focused. The data of PM 2.5 is separated into three periods: Pre-Lockdown, Lockdown, and Phase Opening. Pre-Lockdown covered 41 days from February 29th, 2020 to April 9th, 2020. While lockdown officially established from April 10th, 2020 to June 4th, 2020, from June 5th, 2020 – June 29th, 2020 classified as Phase Opening.

2. Related Studies

2.1. Jakarta’s Air Quality Data Contributors

There are air quality data provider websites that included Jakarta as one of the observed locations. Table 1 compares air quality websites that contained Jakarta as one of the observed areas. Those websites compare four aspects: Record, Values, Access, and Special Feature. In a big picture, the most complete Jakarta’s air quality data parameter ever recorded is PM 2.5. It can be seen by the availability of PM 2.5 on every website. Aqicn.org is the only website that records PM 2.5 and PM 10 in median biweekly while the other two records hourly. On the other hand, openaq.org provides only PM 2.5 while other websites include PM 10 and AQI as well. Among all websites that provide either file download or API (Application Programming Interface), openaq.org facilitates data access with both methods. On top of that, openaq.org allows users to traceback up to 90 days backward for acquiring the desired time range of data while iqair.com stored assorted data that can only be accessed with a paid API. Different from the other two, aqicn.com releases annual files starting from 2015 until today with air quality records from many major cities in the world.

| Website      | Record Format | Values Recorded | Data Access         | Special Feature                                      |
|--------------|---------------|-----------------|---------------------|------------------------------------------------------|
| openaq.org   | Per-hour      | PM 2.5          | File Download, API  | Able to traceback up to 90 days                     |
| iqair.com    | Per-hour      | AQI, PM 2.5, PM 10 | Paid API           | Stored various air quality data from many major cities |
| aqicn.org    | Biweekly      | PM 2.5, PM 10   | File Download      | Releases annual compilation of various air quality data from many major cities |

Table 1. Jakarta Air Quality Data Provider Websites
2.2. Air Quality Analysis

Virgianto and Akbar studied changes in Jakarta’s air quality in Jakarta that occurred because of the 2018 Asian Games [22]. The study discovered that PM 2.5 concentration after the event in South Jakarta was the least (20.2 µg/m³) compare to the condition during and before the event (21.6 µg/m³ and 26.3 µg/m³). The result was obtained by observing daily records from The United States Diplomatic Post at South and Central Jakarta. Descriptive statistics used to get average, median, standard deviation, and maximum and minimum values.

On the other side, Hananto and Putra designed and developed an end-to-end pipeline in the form of a dashboard system to manage and monitor PM 2.5 data in Surabaya [23]. The system supported by IT infrastructures such as a portable sensor device named Edimax Airbox and a cloud platform to store the data in the form of JSON. Utilizing Edimax Airbox enables recording PM 2.5, air humidity, and temperature at the same time and shares the data to the Edimax community website, while the cloud platform facilitates the data easily parsed in the web server and visualized in the dashboard application with various indicators including air pollution index.

The determination of PM 2.5 status change and its forecasting method has been studied by Caraka et al. The study found that the Markov chain model is capable to determine the probability of PM 2.5 that commonly keeps changing among three conditions: no risk, risk, and moderate. The study also found that the combination of Vector Autoregressive, Neural Network, and Particle Swarm Optimization can forecast PM 2.5 levels for the next 180 days. The study was occurred at Pingtung and Chaozhou and obtained the mean absolute percentage error (MAPE) 3.57% for PM2.5 data in Pingtung and 4.87% for PM2.5 data in Chaozhou [24].

3. Research Methodology

The proposed pipeline covered three main steps to produce sufficient data for air quality analysis. There are data acquisition, data pre-processing, and data visualization that defined as core activities of the pipeline.

3.1. Data Acquisition

Database is necessary to contain downloaded datasets and transform them to be an extensive time-series database. Python library SQLite [25] facilitated this research in creating a database, define tables, manipulate rows, execute queries, and manage the database. SQLite has the same workflow with regular MySQL software on the localhost. First, a connection with the .db file must be created. Secondly, a ‘create table’ syntax must be defined to develop the desired arrangement of the database. To import a dataframe, simply add a ‘dataframe.to_sql’ syntax. For special cases like developing an extensive database, a configuration that replacing duplicates must be defined too.

3.2. Data Pre-processing

Python programming language and its fundamental libraries called Pandas and NumPy played a major role in delivering specific treatments [26] and practical solutions in data pre-processing [27]. Figure 1 illustrates a series of data pre-processing activities with Pandas to reshape the raw data into the desired format.

![Figure 1. Data Pre-processing Workflow](image)

To reshape the raw data into the desired condition, six steps are required to perform sequentially. It starts with loading the raw data from the pre-developed database into a dataframe. Since South Jakarta is the focus of this study, the following step is selecting rows containing a certain value in the ‘location’ column. Formats of each column must be examined precisely; a special treatment must be run to make
sure values in the columns contain timestamp-like is containing the desired timestamp format that separates date and time. To simplify data separation based on period, each entry must be labelled according to important dates that officially established by the government of Jakarta. In the second last step, missing data must be detected and replaced by NaN (Not a number) so it won’t be included in the computation. Lastly, the average amount of PM 2.5 is grouped by the hour and periods. This series of data pre-processing activities will produce a time-series dataset explaining the trend of PM 2.5 during pre-lockdown, lockdown, and phase opening.

3.3. Data Visualization
It is important to convey complicated data in a more appealing form through data visualization to make people understand the story lies within comprehensively [28]. Python data visualization library named Matplotlib [29, 30] can transform data into various 2D visualizations. Line plot, scatter plot, box plot, bar chart, and pie chart are examples of data visualizations covered by Matplotlib. In creating data visualization with Matplotlib, one must define columns attached to X and Y axes. To customize the plot, Matplotlib provides various adjustable configurations namely colors, marker, and line style.

4. Results and Discussion
In this section, the result of the data engineering pipeline to analyse Jakarta’s air quality is presented phase by phase. All three major steps of the proposed method were run to analyse PM 2.5 fluctuation at South Jakarta from February 29th – June 29th, 2020.

4.1. Data Acquisition
Openaq.org was chosen as the resource considering the ease of access and the availability of hourly records. To acquire PM 2.5 data on the specific range, the data acquisition phase was run twice. The first download was run on May 28th, 2020 and acquired PM 2.5 from February 29th, 2020 in two locations: The United States Diplomatic Post Central Jakarta and South Jakarta. The second download was run on July 21st, 2020, and acquired PM 2.5 starts from April 23rd at the same locations. Since overlapping dates are found on both datasets, SQLite configured to merge both datasets and drop duplicate rows at once if duplicate dates exist. In the end, SQLite produced a database file with a table containing 9,053 rows from 121 days.

Table 2 describes samples of PM 2.5 data that were successfully acquired into the database system. From the downloaded dataset from openaq.org, four columns are transferred namely: location, date, and time in coordinated universal time (UTC), date and time in Jakarta time (Local), and the amount of PM 2.5 itself that measured in µg/m³. A column named Record ID is generated to order the records incrementally.

| Record ID | Location                  | UTC            | Local                     | PM 2.5 (µg/m³) |
|-----------|---------------------------|----------------|---------------------------|----------------|
| 6623      | US Diplomatic Post:       | 2020-05-24     | 2020-05-24                | 24             |
|           | Jakarta South             | T10:00:00.0000Z| T17:00:00+07:00           |                |
| 3239      | US Diplomatic Post:       | 2020-04-18     | 2020-04-19                | 96             |
|           | Jakarta South             | T19:00:00.0000Z| T02:00:00+07:00           |                |
| 7067      | US Diplomatic Post:       | 2020-05-29     | 2020-05-29                | 19             |
|           | Jakarta Central           | T09:00:00.0000Z| T16:00:00+07:00           |                |

4.2. Data Preprocessing
Table 3 describes the pre-processed PM 2.5 data that has been reshaped into average per-hour and grouped by three different periods. Each period has 24 rows which represent the average of PM 2.5 grouped by hours, from 00:00 from midnight to 23:00 in the evening. Some values are bold and tagged with (A) which represented Apex and (N) which represented Nadir. During pre-lockdown, PM 2.5 starts
at 00:00 in the amount of 45.175 µg/m³ and keep increasing to the apex at 08:00 in the amount of 50.8 µg/m³. The amount decreases to the nadir at 17:00 in the amount of 30.121 µg/m³ and increasing again afterward. When lockdown established, PM 2.5 starts in the amount of 56.264 µg/m³ at 00:00 and increases to the apex at 04:00 in the amount of 63.074 µg/m³. The amount decreases to the nadir at 17:00 too in the amount of 36.759 µg/m³ before increasing back starting from 18:00 until 23:00. In phase opening, PM 2.5 starts at 73.208 µg/m³ and increases to the apex at 02:00 in the amount of 79.041 µg/m³. PM 2.5 decreases to the nadir at 16:00 in the amount of 47.125 µg/m³.

Table 3. Pre-processed PM 2.5 data

| Hour   | Pre-Lockdown | Lockdown | Phase Opening |
|--------|--------------|----------|---------------|
| 00:00  | 45.175000    | 56.264151| 73.208333     |
| 01:00  | 45.700000    | 58.660377| 79.041667     |
| 02:00  | 47.384615    | 60.339623| 81.083333 (A) |
| 03:00  | 50.025000    | 62.735849| 78.375000     |
| 04:00  | 49.600000    | 63.074074| 75.916667     |
| 05:00  | 47.875000    | 61.888889| 70.125000     |
| 06:00  | 46.475000    | 61.148148| 68.750000     |
| 07:00  | 49.225000    | 60.981481| 67.708333     |
| 08:00  | 50.800000 (A)| 58.148148| 68.750000     |
| 09:00  | 44.974359    | 49.698113| 62.500000     |
| 10:00  | 41.200000    | 46.037037| 61.791667     |
| 11:00  | 40.250000    | 45.444444| 57.458333     |
| 12:00  | 41.125000    | 45.203704| 56.000000     |
| 13:00  | 36.902439    | 43.207547| 55.875000     |
| 14:00  | 35.525000    | 39.679245| 55.666667     |
| 15:00  | 32.292683    | 37.129630| 50.958333     |
| 16:00  | 30.268293    | 36.796296| 47.125000 (N)|
| 17:00  | 30.121951 (N)| 36.759259 (N)| 47.875000    |
| 18:00  | 33.121951    | 39.666667| 51.625000     |
| 19:00  | 37.268293    | 41.056604| 55.750000     |
| 20:00  | 38.609756    | 42.471698| 61.458333     |
| 21:00  | 41.926829    | 45.886792| 66.833333     |
| 22:00  | 43.585366    | 49.566038| 72.083333     |
| 23:00  | 43.682927    | 52.754717| 74.708333     |

Figure 2. Visualization of PM 2.5 fluctuation on every periods
4.3. Data Visualization

Figure 2 illustrates the fluctuation of PM 2.5 which previously pre-processed and separated into three periods. The figure compiles three curves that represent the average per-hour values of PM 2.5 on Pre-Lockdown, Lockdown, and Phase Opening. As presented, pre-lockdown’s PM 2.5 starts in a stable amount of 45 µg/m³ at 00:00 – 01:00 and increase until reaches 50 µg/m³ twice. The first peak happens at 03:00 and the second one happens at 08:00. The amount of PM 2.5 gradually decreases to the stagnant lowest point at 16:00 – 17:00 in the amount of 30 µg/m³. Afterward, the amount increases consistently starting from 18:00 until 23:00. In the periods of lockdown and phase opening, the same pattern occurs. During the lockdown, PM 2.5 starts in the amount of 56.264 µg/m³ and climbs to apex at 04:00 in the amount of 63.074 µg/m³. It decreases until meeting a plateau at 11:00 – 12:00 in the amount of 45 µg/m³ and continue decreasing to the stagnant lowest point in the amount of 36 µg/m³ at 16:00 – 17:00. Finally, the concentration of PM 2.5 back to increase from 18:00 until 23:00. In phase opening, PM 2.5 reaches the apex at 02:00 in the amount of 81.083 µg/m³ after starts in the amount of 73.208 µg/m³. The curve declines gradually until meet a plateau at 13:00 – 14:00 in the amount of 55 µg/m³. It decreases to nadir at 16:00 in the amount of 47.125 µg/m³ and continue increasing until 23:00.

4.4. Discussion

Based on table 3 and figure 2, it is visible that PM 2.5 fluctuates on the same pattern even though it occurs in three different lockdown phases. The anomaly is the apex of PM 2.5 occurs earlier during lockdown (04:00) and phase opening (02:00) rather than when it was normal or pre-lockdown (08:00) even though nadir of PM2.5 still occurs at the same time (16:00 – 17:00). If the anomaly related to reference [13] which states that PM 2.5 is contrary to several variables, it can be said that the temperature of South Jakarta experiences a time shift when the temperature decreases to the lowest point.

In Virgianto and Akbar’s study, descriptive statistics were performed to get average, median, standard deviation, and maximum and minimum values of Jakarta’s air quality. Although utilizing the same data source, the amount of acquired data was much less than this study has. The acquired observation data is 63 days for the Asian Games-related study, while 121 days stored in the database for this study. Both of the studies labeled the whole acquired dataset into three categories similarly: during the event, during-event, and post-event.

The research from Hananto and Putra delivered a data engineering pipeline embedded in a dashboard for monitoring air quality in Surabaya with combinations of modern IT infrastructure and devices. This improvement delivered not only a manageable data inventory but also an independent air quality monitoring dashboard for Surabaya. Hananto and Putra’s pipeline is not depending on data contributors such as The United States Diplomatic Post or BMKG (Indonesia’s Agency of Meteorology, Climatology, Geophysics).

5. Conclusion

This paper proposes a data engineering pipeline to acquire time-series data that will continue to grow. The pipeline was run to analyse Jakarta’s air quality from PM 2.5 data from February 29th to June 29th, 2020. The pipeline successfully created an extensive time-series database that leads to a pattern discovery of PM 2.5 concentration at South Jakarta where PM 2.5 drops to the stagnant lowest point at 16:00 to 17:00 although apex reached at different times.

Acknowledgement

We would like to thank Tjeng Wawan Cenggoro and Arif Budiarto of Bina Nusantara University’s Bioinformatics and Data Science Research Center for providing insights that assisted analysis of this research.

References

[1] Detikcom journalist team (2020) Masa Transisi PSBB, Rumah Ibadah di DKI Dibuka 5 Juni. In: detikNews. https://news.detik.com/berita/d-5040265/ masa-transisi-psbb-rumah-ibadah-di-dki-dibuka-5-juni. Accessed 19 Jul 2020
[2] Caraka RE, Lee Y, Kurniawan R, Herliansyah R, Kaban PA, Nasution BI, Gio PU, Chen RC,
Toharudin T, Pardamean B (2020) Impact of COVID-19 large scale restriction on environment and economy in Indonesia. Glob J Environ Sci Manag 6:65–84

[3] Umaseshi RA (2020) PSBB Transisi Dipperpanjang dan Sederet Hal yang Harus Diketahui. In: Kompas.com. https://megapolitan.kompas.com/read/2020/07/02/10043291/psbb-transisi-dipperpanjang-dan-sederet-hal-yang-harus-diketahui. Accessed 12 Jul 2020

[4] Alfons M (2020) PSBB Transisi Mulai Berlaku, Disdik DKI: Sekolah Masih Ditutup. In: detikNews. https://news.detik.com/berita/d-5045902/psbb-transisi-mulai-berlaku-disdik-dki-sekolah-masih-ditutup. Accessed 19 Jul 2020

[5] Sancaya R (2020) PSBB Transisi, Warga Kembali Mulai Berkantor. In: detikNews. https://news.detik.com/foto-news/d-5044737/psbb-transisi-warga-kembali-mulai-berkantor/. Accessed 19 Jul 2020

[6] Sari N (2020) 10 Hari PSBB Transisi, Ada 1.263 Kasus Baru Covid-19 di Jakarta. In: Kompas.com. https://megapolitan.kompas.com/read/2020/06/14/17462841/10-hari-psbb-transisi-ada-1263-kasus-baru-covid-19-di-jakarta?page=all. Accessed 19 Jul 2020

[7] Muljo HH, Perbangsa AS, Pardamean B (2019) Assessment of online learning application for health education. Int J online Biomed Eng 15:69–80

[8] Muljo HH, Perbangsa AS, Pardamean B (2018) Online learning prototype for higher education. Proc 2017 Int Conf Inf Manag Technol ICIMTech 2018 2017:49–53

[9] Pardamean B, Muljo HH, Suparyanto T, Masli E, Prabowo H, Donovan JD (2017) Team Based Learning as an Instructional Strategy: A Comparative Study. New Educ Rev. https://doi.org/10.15804/ter.2017.50.4.11

[10] Rahutomo R, Purwandari K, Sigalingging, Join WC, Pardamean B (2020) Improvement of Jakarta ’s Air Quality during Large Scale Social Restriction. Conf. Second Int. Conf. Biosph. Harmon.

[11] Manzano-León N, Serrano-Lomelin J, Sánchez BN, et al (2016) TNFα and IL-6 responses to particulate matter in vitro: Variation according to PM size, season, and polycyclic aromatic hydrocarbon and soil content. Environ Health Perspect 124:406–412

[12] Huang T, Chen J, Zhao W, Cheng J, Cheng S (2016) Seasonal variations and correlation analysis of water-soluble inorganic ions in PM2.5 in Wuhan, 2013. Atmosphere (Basel). https://doi.org/10.3390/atmos7040049

[13] Yang Q, Yuan Q, Li T, Shen H, Zhang L (2017) The relationships between PM2.5 and meteorological factors in China: Seasonal and regional variations. Int J Environ Res Public Health. https://doi.org/10.3390/ijerph14121510

[14] Kacamarga MF, Pardamean B (2015) Lightweight Virtualization in Cloud Computing for Research. Commun Comput Inf Sci. https://doi.org/10.1007/978-3-662-46742-8

[15] Kacamarga MF, Budiarto A, Pardamean B (2020) A platform for electronic health record sharing in environments with scarce resource using cloud computing. Int J online Biomed Eng 16:63–76

[16] Caraka RE, Chen RC, Toharudin T, Pardamean B, Bakar SA, Yasin H (2019) Ramadhan short-term electric load: A hybrid model of cycle spinning wavelet and group method data handling (CSW-GMDH). IAENG Int. J. Comput. Sci. 46:

[17] Caraka RE, Bakar SA, Pardamean B, Budiarto A (2018) Hybrid support vector regression in electric load during national holiday season. Proc - 2017 Int Conf Innov Creat Inf Technol Comput Intell IoT, ICITech 2017 2018:January:1–6

[18] Caraka RE, Tahmid M, Putra RM, Iskandar A, Mauludin MA, Hermansah, Goldameir NE, Rohayani H, Pardamean B (2018) Analysis of plant pattern using water balance and cimogram based on oldeman climate type. IOP Conf Ser Earth Environ Sci. https://doi.org/10.1088/1755-1315/195/1/012001

[19] Rahutomo R, Perbangsa AS, Soeparno H, Pardamean B (2019) Embedding Model Design for Producing Book Recommendation. Proc 2019 Int Conf Inf Manag Technol ICIMTech 2019 537–541

[20] Rahutomo R, Samsinga A, Wawan T, Pardamean B Artificial Intelligence Model Implementation in Web-Based Application for Pineapple Object Counting.
[21] Budiarto A, Kacamarga MF, Suparyanto T, Purnamasari S, Caraka RE, Muljo HH, Pardamean B (2018) SMARTD Web-Based Monitoring and Evaluation System. In: 2018 Indones. Assoc. Pattern Recognit. Int. Conf. IEEE, pp 172–176

[22] Virgianto RH, Akbar D (2019) Analisis Konsentrasi PM 2.5 Selama Penyelenggaraan Asian Games ke-18 di Jakarta. STATMAT (Jurnal Stat dan Mat 1:44–62

[23] Hananto VR, Putra IGNAW (2018) A Dashboard System for Monitoring Air Pollution in Surabaya based on PM2.5. J Inf Syst Eng Bus Intell 4:139

[24] Caraka RE, Chen RC, Toharudin T, Pardamean B, Yasin H, Wu SH (2019) Prediction of status particulate matter 2.5 using state Markov chain stochastic process and hybrid VAR-NN-PSO. IEEE Access 7:161654–161665

[25] Hipp R (2009) SQLite.

[26] Rahutomo R Preprocessing Methods and Tools in Modelling Japanese for Text Classification.

[27] Albon C (2018) Python Machine Learning Cookbook, First Edit. O’Reilly Media, Inc., Sebastopol

[28] Otten JJ, Cheng K, Drewnowski A (2015) Infographics and public policy: Using data visualization to convey complex information. Health Aff 34:1901–1907

[29] Barrett P, Hunter J, Miller JT, Hsu J-C, Greenfield P (2005) matplotlib -- A Portable Python Plotting Package. ASP Conf Ser 347:91

[30] Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:99–104