International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Validating the Data Acquisition and Serialization for Pollution Data using Big Data Analytics

Subashree D[1], Dhananjay Narayan[2], Alok Kumar[3], Ayush Sharma[4]

[1]*Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*
[2,3,4]*Student, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India*

Abstract - *In this digital era, data is generated in great volume, variety and velocity. Not all of the data generated has significance. Insignificant and redundant data must be eliminated to form a quality dataset. This data generated in terabytes and petabytes leads to the coining of the term Big Data. The data must be optimally analyzed to enable better models to provide high precision recommendations and solutions to serve mankind. A variety of big data tools are employed to facilitate the faster processing of big data. However, there is no enough evidence to prove if the same tools and methods can be used to improve the analysis for relatively much smaller data. To test this, some of the big data methods and techniques are experimented on pollution data to improve the analysis of small data using big data analytical methods. The effect of quality of air on pollution is analyzed. Poor Air Quality is one of the major challenges that a country faces and is one of the leading causes of deaths. We analyse the major constituents of air that causes contamination of air.*

Key Words: **big data, small data, data compression, data serialization, data analysis, pollution, air quality**

1. INTRODUCTION

With the exponentially growing torrents of data, data seems to be generated at a proliferating rate. The main source for big data is generated from the ever-increasing use of mobiles, apps and social media. The data generated by the industrial platforms is yet another source for rich data sets [1]. Big Data coupled with cloud computing leads to scalable data analysis to tell us new things about the world. Big Data provides with better models that allow high precision solutions to make the world a better place. With the advent of smart cities increasingly becoming popular, a plethora of sensors is placed across the geographical locations to read and collect real-world data. The data collected by these sensors prove to be a significant contribution to Big Data. These sensors are effectively used to collect the pollution data to find suitable techniques to minimize the damages caused by pollution and efficiently control it. However, not all of the data collected is useful. This data must be cleaned and validated. Data Analysis must be performed to make sense of the data collected. Optimal data acquisition, compression and serialization are pivotal processes in data analysis [1].

The data acquisition process deals with the collection of data and converting it into numeric values such that it can be manipulated by a computer. Data Serialization involves the conversion of structured data in a format that allows the recovery of its original structure. The main use of data serialization is to reduce the size of the data which further minimizes the disk space and bandwidth requirements.

One of the misconceptions of Big Data Analytics is that you need data in the size of terabytes or petabytes [2]. Though the pollution data generated is not as rapid and vast as other big data sources such as social media and industrial data, methods for optimal data analysis is equally important. This paper demonstrates the processing of small data. The small data in this instance is of pollution. We use the scope of data analytics to determine the effects of factors such as the quality of air on pollution.

2. EXISTING SYSTEM

There are various data engines that perform the big data processes such as compression, serialization and analysis. One such platform is HortonWorks Data Platform (HDP), which integrates many of the big data technologies. It uses distributed storage and processes large data sets to enables agile application deployment [3].

Another design called iKayak runs Spark and MapReduce on YARN clusters, which dynamically optimizes resource scheduling for big data. [4]. HANA platform was designed to inhibit the networking features for Big Data, which helps in gathering the rapidly generated data and moves them to data centres. The network designs were not extendible to interconnect multiple data centres and server nodes [5]. Another Big Data Acquisition and Storage System (ASS) for industrial data platform showed improved performance in data compression using LZ0 algorithm. Protobuf proved to improve the serialization process [1].

While many methods were proposed for richer big data sets, there is no enough evidence to verify if the same big data techniques have the efficacy to improve the processing for small data sets. We use some of the big data tools and techniques on pollution data to analyse it.

3. PROPOSED SYSTEM

As not lot of big data methods has been tested on the smaller data sets, there is no evidence to demonstrate how the same techniques will work on them. We have chosen pollution data set as sample and will test the performance of big data methods on this collected dataset. The data sample collected is cleaned and validated. This cleaned data is then loaded and stored. A set of compression algorithms and serialization techniques are employed on the stored data. The data is then visualized to analyse the air quality to determine the significant factors that contribute to the pollution. Air quality analysis is done to make sense of the data collected and find significant pattern. The data is visualized according to the proper attributes to create charts and conclusions are drawn based on these charts.

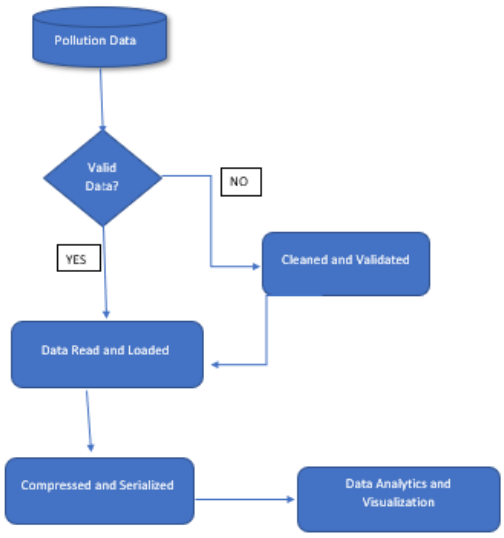

Fig -1: Work Flow Diagram for Proposes System

4. SYSTEM ARCHITECTURE

The diagram in Fig 2 is a complete representation of how the system functions and the various stages and components involved. We are taking pollution data set as sample here. The data sample collected is cleaned and validated. This cleaned data is then loaded and read using Pandas. The data is compressed and serialized. The data is then visualized to analyse the air quality to determine the significant factors that contribute to the pollution.

First, the selected pollution dataset of a particular state or a country is chosen and is acquired from the government website. This dataset is partially cleaned. However, we still need to check if it is completely clean. We first check if the data in the CSV file is valid or not. If the data is not valid, it is cleaned and then validated. Validation of data refers to verify if the data is in correct format or not. If the data format is invalid, it can lead to various errors during the analysis and can lead to the unreliable results, which can be fatal in the real-world. In data cleaning we clean the raw data by removing the null values and unnecessary columns. Hence, Cleaning and Validation of Data is of utmost importance.

After the data is cleaned, we then compress and serialize the data. The main objective of compression is to save the storage space. Compressing the data also helps in the reduction of bandwidth required for data sets. It can also remove redundant data, which makes the analysis faster and easier. Serialization helps in converting data in a format such that it allows easy recovery of its original structure. The data can be structured as either flat or nested. For flat data, the CSV module of python allows classes in reading and writing table data in CSV format.

Once the compression and serialization of the data is done, it is now time to perform the data analysis. Data analysis is a process where we collect and organize data in order to draw useful conclusions from it. The main use of doing it is to find

meaning in data so that it can be used for making informed decisions. There are various data analysis tools that makes it easy to process and manipulate data. These tools help in finding correlations in the data sets and helps in identification of patterns in data. Some of the data analysis tools are SQL, SAS, MATLAB and python libraries. In our project, we make use of some python libraries to analyze data. Data visualization is a method to present the graphical representation of data. Charts, graphs and maps can be created by the visualization tools which provided an accessible way to understand the trends and patterns in data. This plays a key role to make data-driven decisions.

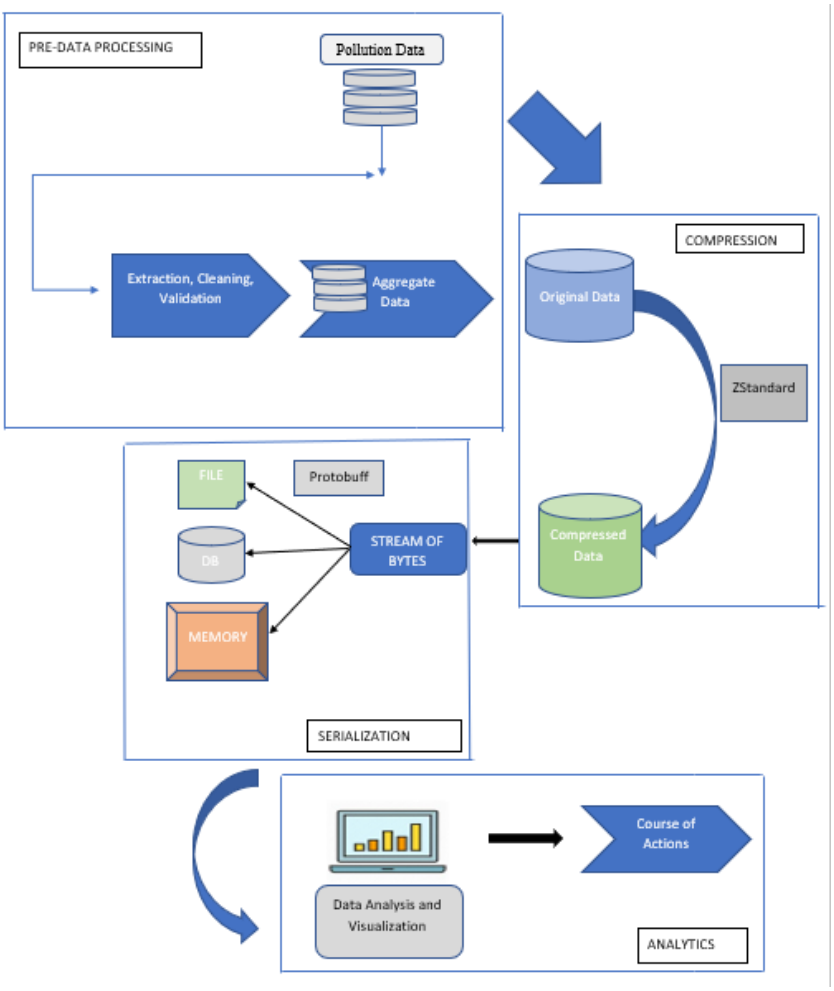

Fig -2: System Architecture

5. MODULARIZATION

5.1 Data Pre-Processing

Preprocessing of data is a technique used in the data mining field whose purpose is to convert the raw data provided employing dataset into a useful and efficient format which will help to easily remove the unwanted or noisy data from the dataset. Preprocessing has three parts which are data cleaning, data transformation and data reduction. Partially cleaned data is acquired from the repositories. The extracted data is further cleaned. They are briefed below –

1. Data Cleaning- We clean the raw data by removing the unnecessary columns and the null data. In missing data, we can ignore that cell or fill it manually with the most probable values.

2. Data Transformation- Data transformation is carried out in order to convert or transform the data in suitable forms for the data mining process. It includes various methods such as normalization, attribute selection, discretization and concept hierarchy generation. The columns can be renamed and transformed as needed.

3. Data reduction- In data mining or big data we use large volumes of data because of which the analysis part becomes harder. To solve this problem, we use reduction technique of data, which utmost purpose is to reduce data storage and the analysis costs.

```
In [7]: del ds['Stn Code']

In [8]: del ds['Location of Monitoring Station']

In [9]: ds.head() # To check that the said columns are deleted
```

Out[9]:

	Sampling Date	State	City/Town/Village/Area	Type of Location	SO2	NO2	RSPM/PM10	PM 2.5
0	03-01-2015	Karnataka	Mysore	Households	10.0	22.0	39.0	NaN
1	06-01-2015	Karnataka	Mysore	Households	11.0	22.0	40.0	NaN
2	09-01-2015	Karnataka	Mysore	Households	11.0	22.0	37.0	NaN
3	13-01-2015	Karnataka	Mysore	Households	11.0	24.0	44.0	NaN
4	17-01-2015	Karnataka	Mysore	Households	11.0	22.0	48.0	NaN

```
In [10]: ds.isnull().sum() #to check the count of null values
```

Out[10]:
```
Sampling Date                1
State                        1
City/Town/Village/Area       1
Type of Location             1
SO2                         41
NO2                         41
RSPM/PM10                    1
PM 2.5                    2393
dtype: int64
```

```
In [11]: ds = ds.dropna(axis = 0, subset = ['Type of Location'])

In [12]: ds = ds.dropna(axis = 0, subset = ['SO2'])

In [13]: ds = ds.dropna(axis = 0, subset = ['NO2'])

In [ ]: del ds['PM 2.5'] # Too many Null Values. So deleted the entire column.
```

Fig -3: Data Pre-Processing Module

5.2 Compression

Compression of data is a process of encoding, modifying, and reducing the numbers of bits needed for the data representation. Compression of data can help in saving the storage capacity, transfers of files get the speed up and the cost for storing the data also gets minimized. It is done performed by using a various algorithm which determines how to shrink the size of the data. The algorithm may use dictionaries or pointers for conversions and referencing of bits to a smaller one. Data compression reduced the amount of storage a file needs and also optimizes the storage backup performance.

As the data grows exponentially in the big data environment data compression plays a vital role. Real time compression algorithm like ZStandard can be used to compress the data, which can provide high compression ratios. It offers a special mode for small data called Dictionary Compression. Dictionary files are created from sample sets. Using this dictionary, the ratio achieved on small data compression improves [6].

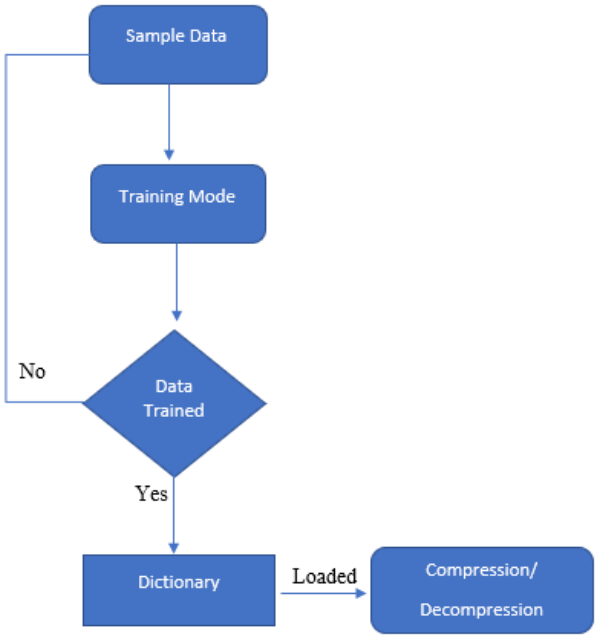

Fig -4: Data Compression Process

5.3 Serialization

Serialization of data is a process in which data objects present in the forms of a complex data structure are converted into bytes of the stream which could be used for storing, transferring, or distribution purposes. Since, the computers vary per their OS, architecture, and mechanism of addressing [7]. Storage and exchange of data in such varying environments is a tough task to perform for which it requires a neutral data format for an understanding of every system present. Data serialization formats depend upon the choice and factors such as speed, data complexity, and the constraints for storage space. BSON, YAML, JSON, XML, and protobuf are some data serialization formats used commonly.

We will take the compressed data after the compression process and which is then used for the serialization process. It will be converted to a stream of bytes which could be used for various purposes such as to store it in the database, file, or memory. We use the serialization formats such as protobuf which produce a major result for the serialization process. The data is then serialized in a way that it could be used on any platform and it doesn't depend on the architecture of its system.

5.4 Data Analytics and Visualization

An analysis is a process in which the data is collected and organized in a way to retrieve useful information from it. It uses logical and analytical reasoning to gain information from the data. The analysis helps in making better decisions, it provides the insights of data to make the right choices. In our data analysis and visualization part, we will be generating a graph of data for the various parameters of the air pollution which are present in that region. The graphs will provide a better understanding of the constituents such as SO2, CO, RSPM in the air which could be harmful to the living organisms. By analyzing all those graphs, we can conclude to take the actions which would be useful in saving that area from the pollution of air which defines our course of action.

6. EXPERIMENTAL RESULTS

The dataset is taken of Karnataka. There are around ten attributes present in the dataset. Not all these attributes are important in analyzing the air pollution. The irrelevant columns such as Station Code, Location of Monitoring and the Agency are removed which further saves some space as well. The null values are checked and eliminated. The column 'PM 2.5' had the most null values. These along with other null values were eliminated. The concentrations of SO2, NO2, and RSPM were plotted in graphs.

The scatter plot for each of the columns was created. A correlation matrix graph was visualized to see how correlated the data were. After adding a new column for year, heatmaps were plotted for each SO2, NO2 and RSPM.

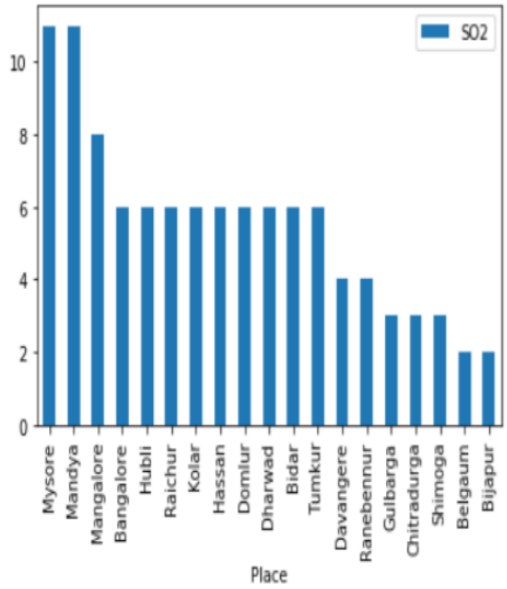

Fig -5: SO2 Concentrations in Cities

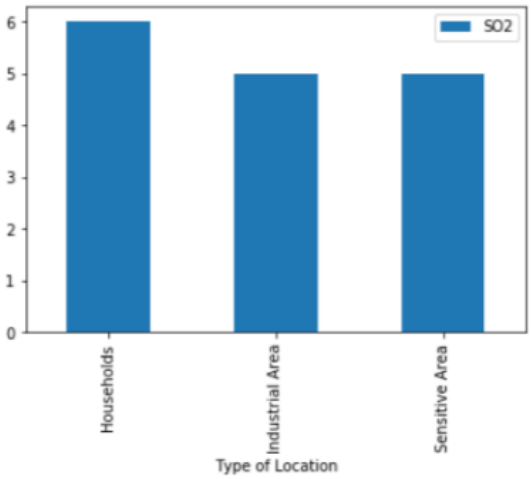

Fig -6: SO2 Concentrations vs Type of Location

From graph in Fig 5, we can deduce the finding that SO2 intensification is at the peak in the cities of Mysore and Mandya. In the other places, there is equal distribution of SO2 concentration. It is the lowest in Belgaum and Bijapur (now Vijayapur). From this, it will be easy for the government to decide on where to take the further actions so as to control the concentration of SO2 in these cities. Here in Fig 6, we can see that the household localities have the highest SO2 levels while the concentrations are almost comparable in the industrial and the sensitive areas.

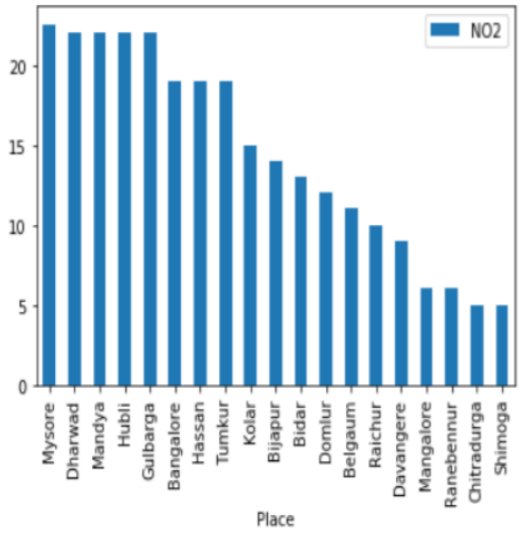

Fig -7: NO2 Concentrations in Cities

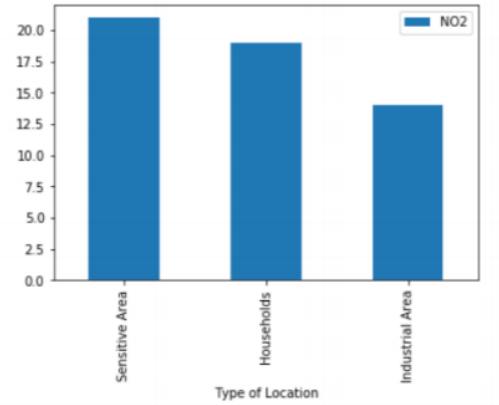

Fig -8: NO2 Concentrations vs Type of Location

It is very clear from the graph in Fig 7 that, Mysore, Dharwad, Mandya, Hubli and Gulbarga has the highest concentration of NO2 with values above 20.0. Bangalore, Haasan and Tumkur follows next with values around 17.0. When we combine both the SO2 and NO2 concentration results, we can conclude that Mysore and Mandya are one of the most populated cities and needs immediate attention and suitable actions on priority to make the air heathy to breath for its citizens. From Fig 8, we see that NO2 is highly concentrated in the Sensitive areas. The industrial areas have the least values. Chitradurga and Shimoga are very clearly having the least values of NO2. These two cities have values just around 5.0 and has comparatively safer air to breathe in when compared to other metropolitan cities in the state of Karnataka.

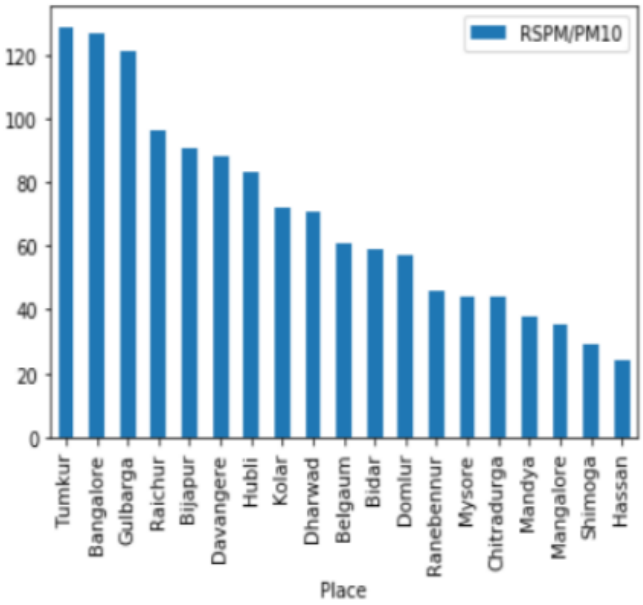

Fig -9: RSPM Concentrations in Cities

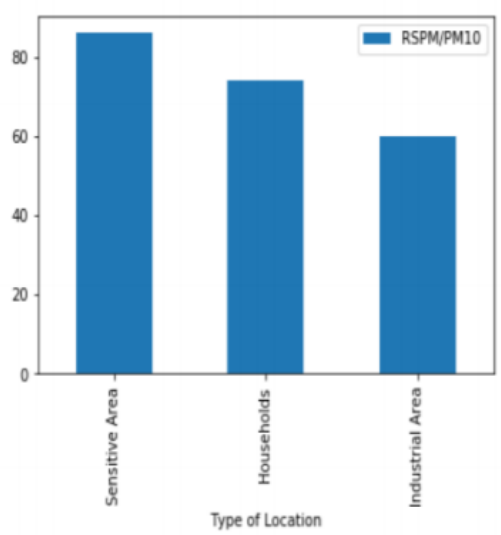

Fig -10: RSPM Concentration vs Type of Location

The figures 9 and 10 gives the concentration values of RSPM. It is the most dangerous pollutant that has very adverse effects if not controlled in time [8]. The cities of Bangalore, Tumkur and Gulbarga have the highest concentrations. Bangalore has consistently been a driving objective with regards to contamination. The people of Bangalore suffer a lot due to this and it has been a cause of deaths in several cases. We also see that the city of Bijapur and Raichur are not far behind. Without much surprise, the major factor in these high values are due to the sensitive areas. Proper deterrents must be in place to further avoid the increase in the concentration levels of RSPM.

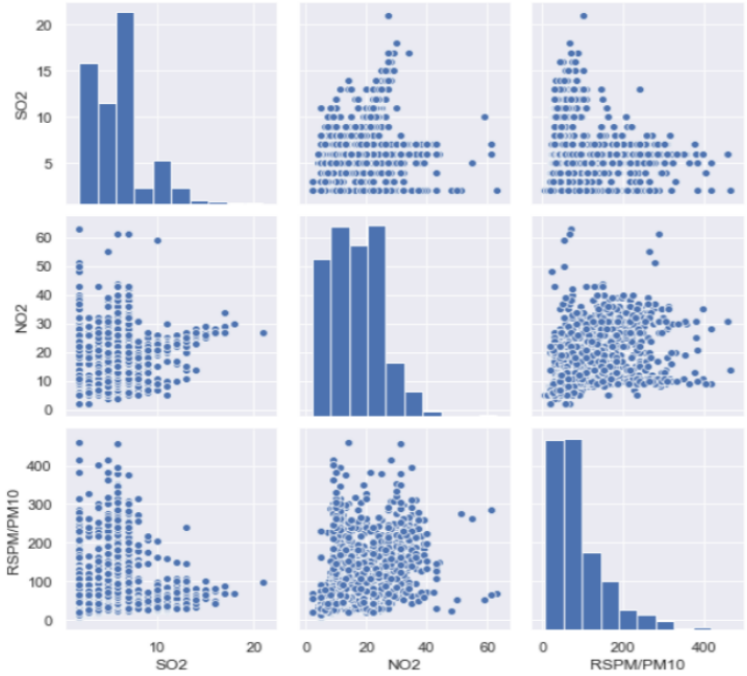

Fig -11: Scatter Plot for Each Column

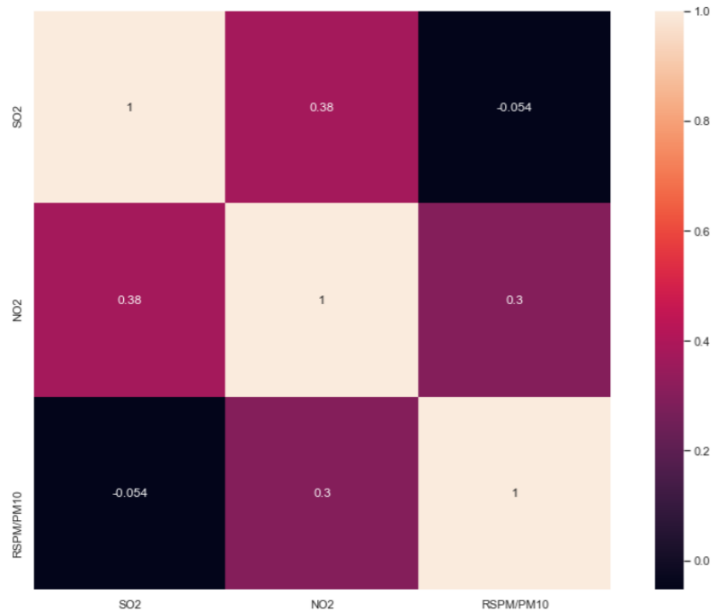

Fig -12: Correlation Matrix for Dataset

We do some statistical analysis in Fig 11. We are checking if these features have some relations. We have plotted a scatter plot for each constituent of pollution from the dataset. Nitrogen Dioxide and Sulphur Dioxide are very highly concentrated in the points of the origin. This suggests that they are actually significantly lower for most of the values observed. It also seems that the attributes in the dataset do not follow a similar pattern. They are mostly not related. To confirm this in depth, we have plotted a Correlation matrix graph in Fig 12. Here, it has become very clear that the values close 0.3 out of 1.0 supports our earlier judgement. We might have deduced that these features have some correlation if the matrix values were close to 1.0. We also see that the correlation between the values of SO2 and RSPM are below zero. This signifies that these two values are not at all dependent on one another. Their significance is definitely independent of one another.

With these observations at our hands, we can now plan for actions to curb these growing levels of concentrations.

7. FUTURE ENHANCEMENTS

Data for the recent years is in process of collection. This new data can be used in the existing project to see the timeline from then to now. We can see which cities are improving over time. We can incorporate other attributes like temperature and humidity and can study how the changes in these values affects our environment. We can also study the relation of temperature changes in the depletion of ozone layer. With the field of machine learning booming, we can train the datasets and can use this model to predict the future trends in the air pollution and can take suitable actions beforehand to prevent the damage.

8. CONCLUSION

The pollution dataset for Karnataka was successfully cleaned, validated and compressed. With removing the null values and irrelevant columns, the size was reduced. From the analysis, it is seen that the urban and suburban places like Bangalore, Mysore, Dharwad and Mandya are hugely affected by air pollution and require immediate action. SO2, NO2 and RSPM are the major contributors towards pollution among which RSPM is far more dangerous. From the heatmap, we were able to see that the cities remained to be heavily populated even at the end of year with in decrease in the concentrations of contaminants. Data analysis, whether used for big data or small data, can play a crucial role in planning for a better future. These data-driven approaches can be used to analyze and solve the daily significant problems.

REFERENCES

[1] Daoqu Geng, Chengyun Zhang, Xue Xia, Xinshuai Fu, Chengzing Kia and Qilin Liu, "Big Data based improved acquisition and storage system for designing industrial data platform," IEEE ., March 2019
[2] https://blog.syncsort.com/2017/11/big-data/5-big-data-myths/
[3] P. Basanta-Val, "An efficient industrial big-data engine," IEEE Trans. Ind. Informat., vol. 14, no. 4, pp. 1361 1369, Apr. 2018.
[4] D.Cheng, X. Zhou, P.Lama, J.Wu,andC.Jiang,"Cross-platform resource scheduling for spark and mapreduce on YARN," IEEE Trans. Comput., vol. 66, no. 8, pp. 1341 1353, Aug. 2017.
[5] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a network highway for big data: Architecture and challenges," IEEE Netw., vol. 28, no. 4, pp. 5 13, Jul. 2014.
[6] https://facebook.github.io/zstd/
[7] https://www.geeksforgeeks.org/serialization-in-java/
[8] https://towardsdatascience.com/india-air-pollution-data-analysis-bd7dbfe93841