*International Journal of Engineering & Technology*, 7 (2.33) (2018) 919-923

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Air pollution analysis using big data technology: towards a better world

**Siva Krishna kvs [1] *, Saikumar Pulluri [1], Kamalakannan J [1]**

*[1] School of Information Technology, Vellore Institute of Technology, Vellore*
**Corresponding author E-mail: venkatasaisiva.krishna2013@vit.ac.in*

**Abstract**

Big data is generally perceived as being one of the most intense drivers to advance profitability, Enhance effectiveness, furthermore, bolsters advancement. It is quite anticipated which would analyze big data and transform big data into big values. To find the answer of the fascinat-ing question whether there are characteristic connections between the two inclinations of big data and green challenges, a study has exam-ined the issues on greening the entire life cycle of big data frameworks. As the data which is captured from different sensors is huge, to analysis that data and find patterns to predict the future data, we need big data technology which can handle that huge amount of data in a better way. In this paper, we have used different classifiers to analysis the results based on available data in the spark framework using the Python and Scala programming languages. We showed a comparative study between python and Scala technology based on classifiers. For this research data set of Andhra-Pradesh and Tamilnadu (states in India) are utilized to show the analysis of air pollution with the help of big data concept. We compared the classifiers on based on time and accuracy. Generally random forest gives good results but in our case deci-sion tree and logistic regression have given high accuracy.

***Keywords***: *Air Pollution; Big Data; Classifiers; Decision Tree; Logistic Regression; Naïve Byes; Pollutants; Random Forest.*

## 1. Introduction

According to WHO report 2016, India is one of the most polluted countries in the world. In the list of the top 20 most polluted cities, 13 are from India, in which Delhi is on top. With all these alarming reports urban air pollution is increasing day by day. All urban communities are reeling under serious particulate pollution [14] while newer toxic like oxides of nitrogen and air toxics have started to add to the general health challenge [15].

In the course of the most recent decade, an expanding number of environmental issues have required the investigation of data from hundreds or thousands of various areas, drawing on provincial, national, and worldwide monitoring networks. These far reaching data sets can give essential setting to contrasting current with verifiable conditions at a particular site [14]. Air pollution, the "site" is a whole district, country, or the world, requiring an all-encompassing appraisal utilizing huge scale data sets. As an after effect of the huge volumes and wide assortment of data sorts accessible now to researchers, arrangements progressively require the tools and methods of "Big Data [4]." Advance in information gathering, counting capacity and availability imply that we have more data than any time in recent memory readily available. IBM predicts that by 2020 there will be 300 times more data on the planet than there was in 2005– An aggregate of 43tn gigabytes. Also, this information is being put to great utilize [6]. Progressively we hear how legitimately understanding information prompts to positive results, whether this is related to sports games result prediction or forecasts of the political elections.

Spark:

Spark is an open source community in big data which handles even petabytes of data very efficiently. Spark is Hadoop based platform which used in data analytics. Spark has cluster manager and distributed storage systems [4]. This cluster manager takes care of other distributed systems. Its compatibility with Hadoop makes it more efficient handling the large data and work on machine learning algorithms.

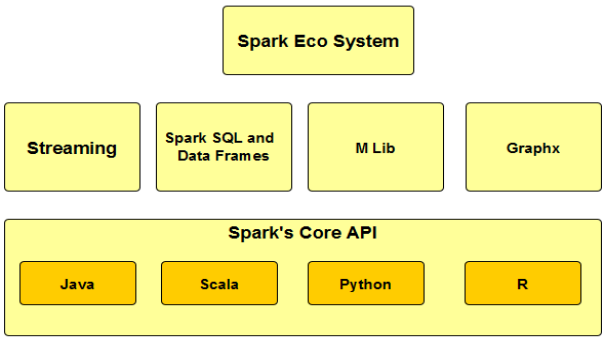


**Fig. 1**: Spark Ecosystem.

In this research work we have presented literature survey in first part, where we discussed previous work details. Second part where we discussed about the architecture and process flow we followed. The infrastructure we required for this project. In third part we discussed the analysis and results we have taken from the execution of data. At last we concluded the results and future enhancement of the work.

## 2. Literature survey

Fan Jiang, Hao Zhang, Carson K. Leung, and Adam G.M. Pazdor have represented a science model for big data analytics [1] for frequent pattern with Map reduce. They used social application

data and real-life applications data to analysis useful pattern from the datasets using the science model and different mining algorithm. There are various frequent pattern mining algorithms have been invented, which include tree-based algorithms, level-wise apriori-based algorithms, and hyperlinked array structure-based algorithms. These algorithms have certain advantages and some of disadvantages.

To mine frequent patterns from huge data, few algorithms have been proposed [1]. For instance, Fixed Passes Combined-counting (FPC), the Single Pass Counting (SPC), and Dynamic Passes Combined-counting (DPC) algorithms [10] were implemented—based on the Apriori and the Count Distribution algorithms—to mine frequent patterns from big precise data using the MapReduce model. A. Cuzzocrea and C. K. Leung has been proposed Parallel Frequent Pattern-growth (PFP) algorithm [10], which also uses the MapReduce model, parallelizes the tree-based FP-growth algorithm for mining frequent patterns from precise big data.

Mihaela Opera, Hai-Ying Liu invented a knowledge-based approach which used for air pollution effect analysis in the case of PM2.5 air pollutant [2] which have some negative effects on human health. They added more rules to make one decision support system. Development of new intelligent tool for atmosphere issues analysis will help us to improve the quality of life using artificial intelligence and machine learning algorithm.

Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner, Daniel Price have done comparison between spark programming and Hadoop MapReduce model [3] for air quality simulation in the state of Texas for a variety of pollutants. This study reveals that spark gives better performance over MapReduce. This study also identified performance benefits of the Spark MLlib machine learning library over the MapReduce Mahout library.

Navjot Kaur Walia, Parul Kalra, Deepti Mehrotra used Naïve Bayes approach to train a model to classify forest data [5] based on the stock of carbon and predict the level of carbon in forest based on the previous available data. This study implies that carbon stock is a noteworthy concern towards tending to environmental change. Since developing countries and increasing pollution impacts the atmosphere so analysis of environmental change is vital later on.

David G. Rickerby, Andreas N. Skouloudis, have presented an idea of cloud applications taking care of enormous information and online networking permits precise continuous geological portrayal of populace presentation [6]. This work concentrates on practical use of photochemical air-quality information from estimations that permits another period for administrative applications in a few land scales.

James Manyika, Michael Chui, Brad Brown generated a report based on big data uses in different fields in real world applications. Examining huge informational collections [8], called big data will turn into a key premise of rivalry, supporting new rushes of efficiency development, advancement, and customer excess.

Nancy and Arushi from IIT Delhi studied that monitoring the air quality to check the air quality level helps in analyze the effects on human being. To check air quality index monitoring station measuring air pollutants like PM10, PM2.5, NO2, SO2 and CO [9]. But according to Nancy and Arushi other than these pollutants benzene, lead, ammonia and few other parameters monitoring also required because these pollutants are more hazardous to health.

They further explained the main sources of benzene and its short term and long-term effect. Other than that, they have given a brief idea how we can measure benzene using sensors and what are the challenges observing the values out of it.

Sreemoyee Roy and Abhik Mukherjee represented a case study on importance of ISA (Information System analysis) for monitoring of air quality. The main problem is distinguished inaccessibility of accurate data in persistent mode [10]. In this work they attempt to distinguish the gaps in this respect and furthermore take a gander at what might be accomplished for a good Air Quality Information System (AQIS) [10] design. They explained the importance of AQIS, the requirements to develop good AQIS and algorithms required to implement the AQIS. They analyzed the Howrah district industry data and compared with the other cities of India, which helped in better analysis of AQIS requirements and challenges.

Lily Bui [11] has presented one study on how the sensing of pollutants happens by monitoring stations. In addition, what challenges they face during the process. He basically gave a clear view on how data collected from sensors and provided on different portals, so it can be accessible to public.

Wenjun Lv, Yu Kang, Zerui Li, Yunbo Zhao have invented a filter named as weighting filter which is based on the traditional filter Kalman filter [12]. It is utilized to estimation for road air contamination focus is preparing each time step, and the estimation is right with accessible estimations. Besides, a self-tuning controller is acquainted with alter the parameters of channel for the changing commotion measurable qualities after some time which principally brought on via season switch. So overall, we can say many efforts have been putting by researchers to reduce the pollution rate in the world from finding a new algorithm to invent a decision support system. They are using artificial intelligence, big data technology, cloud computing, finding new pollutants in air, decision rules, knowledge base system and etc. to get the patterns. Researchers are trying to find the best technology by comparing the results. In this we are comparing python and Scala's results.

## 3. Architecture

This paper mainly contains the analysis of Air Pollution data of Tamilnadu and Andhra-Pradesh. We used the cloud infrastructure to process the data and to store the data we used the S3 AWS (Amazon Web Service) [16].

Amazon Simple Storage Service (S3) provides storage on cloud. It is intended to make web-scale figuring less demanding for developer.

Amazon S3 has a decent web service interface that you can use to store and recuperate any measure of data, at whatever point, from wherever on the web. It gives any developer access to the same exceptionally versatile, strong, speedy, cheap data storage structure that Amazon uses to run its own overall arrangement of sites. The organization arrangements to lift the focal points of scale and to pass those preferences on to developer.

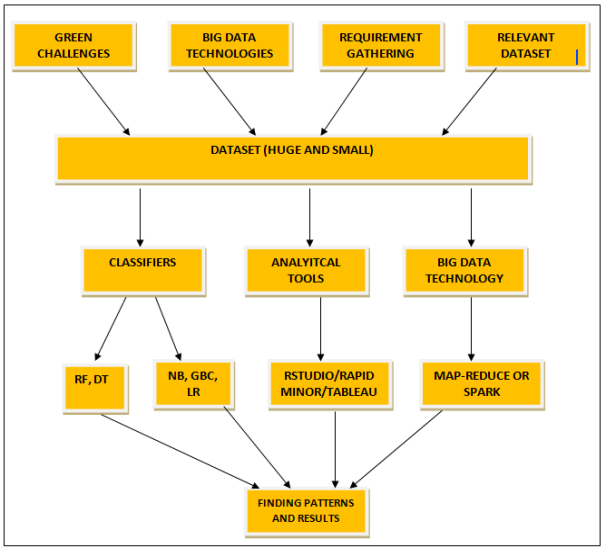


**Fig. 2:** Architecture Flow.

The process flows we followed for this research work is given below. We started from the searching the best suitable big data technologies for green challenges. We are using Spark platform to process the data using the python and Scala languages.

We used few classifiers to analysis how they predict the future data and level of accuracy. We compared the python and Scala programming on spark distributed platform. Python and Scala both used for data science and data analysis project. Both have

their own advantages. Both can be compared basis of productivity, speed, and time taken to run the code, safe refactoring, spark integration, statistics packages, and community and there are several other features.

Other than this we used Tableau tool for analytics and better visualization. This helps in presenting the data in pictorial form.

a) Datasets

There are air pollution monitoring websites which provide the live data, Central Pollution of Control Board (CPCB) and Open Govt Data (OGD) platform India. Real time air quality index is also available for world map based on the air quality index standard.

We have taken dataset from the OGD platform. We have taken six years (2009-2015) data which contains few attributes like SO2, NO2, PPM, Sampling Date, Location, Station Code, Station Name and few more. Based on the standard of SO2, NO2 and PPM standard values, we normalized the data. Normalization of data includes cleaning, handling missing values, finding dependency of attributes and structuring the data. Datasets contains 12352 rows in Tamilnadu's data and AP data contains 15353 rows.

## 4. Method and approaches

Classifiers:

Classifiers are mathematical function or logic which implements classification problems, or we can say an algorithm which classifies the data based on the training and testing data, called classifiers. In below diagram we have shown the working of classifiers. We have to define the ratio of testing and training data before classifying the data.

There are various machine learning algorithms. We have discussed four classifiers Decision tree, Random Forest, Logistic Regression, Naïve Bayes algorithm. We also presented a comparative study among four of them

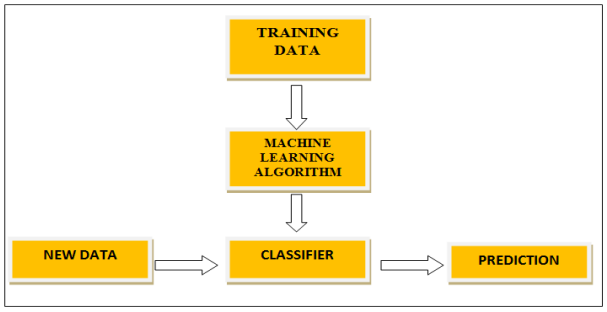


**Fig. 3:** Classifier is working.

a) Decision Tree

Graphs Decision tree is an algorithm for classification and regression. It builds models in a tree structure format. It is easy to understand and easy to implement. It is faster than other classifiers [6]. Its gives high accuracy compared to other classifiers.

Decision tree supports both numerical and categorical data. It cuts down dataset into small and small parts while in the meantime a related decision tree is incrementally made. The last result is a tree with decision nodes and leaf nodes. It is a predictive model which classifies a value based on the present data or classes in leaf's.

b) Random Forest Algorithm.

Random Forest is one of the ensemble algorithms which use more than one learning algorithm for better prediction. There are algorithms which use prediction based on a single tree classification and on other end random forest algorithm which uses multiple classification trees [2]. Random forest is using to averaging the multiple deep decision trees. For the classification of new item from an input vector, we associate a vector point to each tree. Every tress gives a value as a classification or we can vote, based on highest votes we decide the final value for the new data object.

It provides high accuracy. It works well with huge datasets. Good in describing which variables are good in classifications. Created forest data can be used in future for other data prediction. It can be extended to, data views, outlier detection, unlabeled data and leading to unsupervised clustering.

c) Logistic Regression

Logistic regression model can be binary or multiclass. It depends on the outcomes of the analysis [6]. If outcome is binary [0] or 1and true or false, it will be binary ex: good credit people or bad credit people. If outcomes are more than two like our project case, it will be multinomial case. It is mostly used in medical field, other than in machine learning and data science field. Logistic regression defines a best fit logistic function [5]. This probability function use to predict the class of new data based on the previous training data.

d) Naive Bayes Algorithm

Naïve Bayes is a classification algorithm which is based on the Bayes theorem. It follows the assumption of independence of predictor variables [5]. When we have large dataset and les attributes to compare, the best classifier we can use for such problem is naïve Bayes algorithm. When there are multiple classes we can use Naïve Bayes algorithm.

It works well with huge data. It can handle multi class data. It gives good result with categorical data. It is used for real time prediction, spam filtering, sentimental analysis and text classification application.

## 5. Analysis and results

As we analyzed the different classifiers, we recorded the accuracy and time taken by execution. As we can see the time difference is there between python and Scala. We have taken 2MB of data. As the data size is not that large but if with 2MB its showing sec of difference, if there is data in GB or TB, this will make a huge difference.

If we talk about the accuracy, it depends on the parameters we have chosen. It depends on the dependent and independent variables or predictor. For some attributes it will give high accuracy, for some it will give low accuracy. We have taken different variables and checked the accuracy. As the parameters are discussed initially, we have taken few parameters and checked the results. The tables given below shows accuracy and time based on SO2 and Station Code parameters.

**Table 1:** Classifier's Results Analysis (Tamilnadu)

| Classifier Name | Python Execution Time(sec) | Scala Execution Time(sec) | Python Accuracy (%) | Scala Accuracy (%) |
|---|---|---|---|---|
| Decision Tree | 27.00 | 24.00 | 84 | 84 |
| Logistic Regression | 27.00 | 18.00 | 65 | 67 |
| Random Forest | 30.14 | 27.16 | 85 | 86 |
| Naïve Bayes | 17.74 | 16.22 | 53 | 53 |

**Table 2:** Classifier's Results Analysis (Andhra Pradesh)

| Classifier Name | Python Execution Time(sec) | Scala Execution Time(sec) | Python Accuracy (%) | Scala Accuracy (%) |
|---|---|---|---|---|
| Decision Tree | 25.47 | 24.00 | 100 | 100 |
| Logistic Regression | 16.70 | 16.54 | 100 | 100 |
| Random Forest | 30.81 | 28.32 | 100 | 100 |
| Naïve Bayes | 27.50 | 25.00 | 100 | 100 |

For our data decision tree and random forest showed better result than other classifiers. We have taken the training and testing data,

based on trained data predicted the value of testing data. The results we have observed are given in the table 1.

Tableau Tool Analysis:

Other than this we have analysis the data in tableau tool, which gives the clear picture on what days, in which area pollution is high. The level of different toxic gases is presented with the detail of station code.

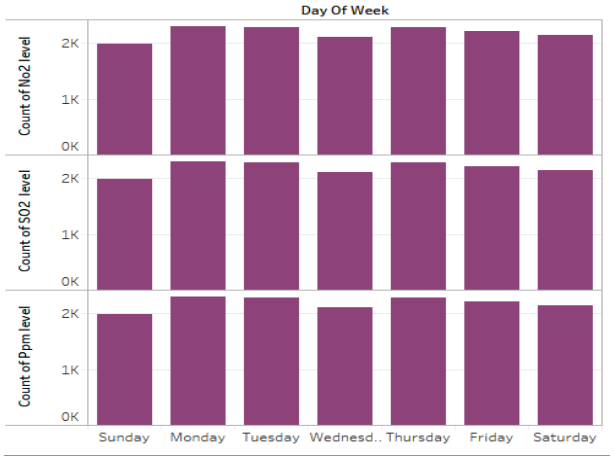


**Fig. 4:** Andhra Pradesh Data According to Weekdays.

As we can see in fig 6 pollution on Monday and Thursday is very high as compared to other days.

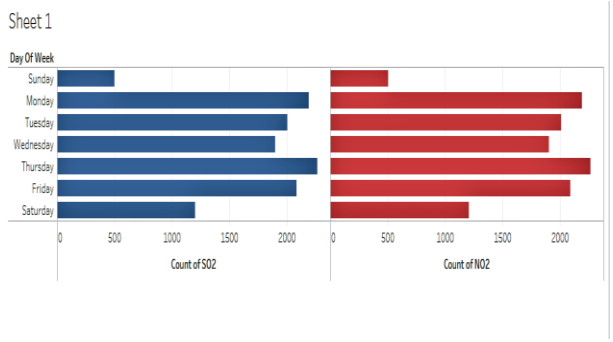


**Fig. 5:** Tamilnadu Data Based on Pollutants and Weekdays.

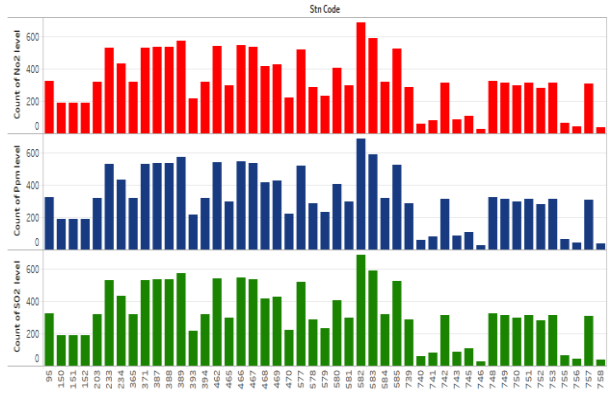


**Fig. 6:** Andhra Pradesh Data Based on Station Code.

Fig 6 represents pollution level of pollutants at different stations. 582 and 583 is having higher degree of pollution than other station.

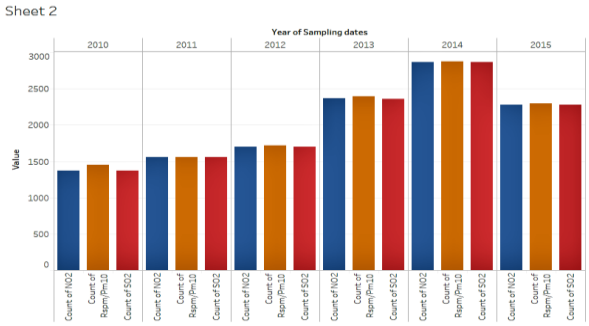


**Fig. 7:** Tamilnadu Data Based on Years.

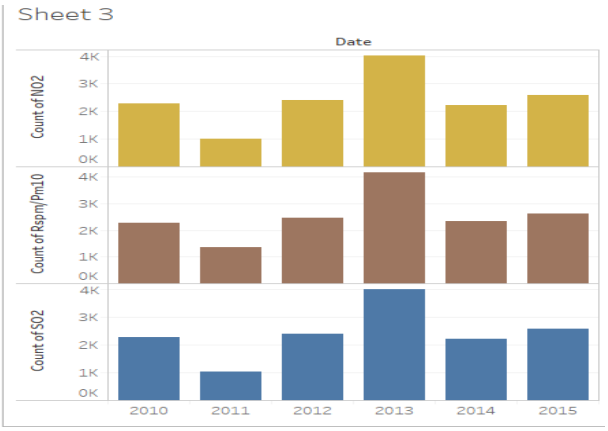


**Fig. 8:** Andhra Pradesh Data Based on Years.

Fig 7 and 8 represents year wise data from (2009-2015) in which Tamilnadu has high level of pollution in 2014 and AP has high level of pollution in 2013 compare to other years.

CPCB Analysis:

Central Pollution Control Board provides a live dashboard where we can analyze the data of different states. We have analyzed the available data on the portal and compare the data we have analyzed. We are presenting few of the analysis work captured at CPCB portal.

Below diagram shows values of pollutants at Alandur Chennai. If we check the graph, value of SO2 has a high variance from year 2013 to 2016.

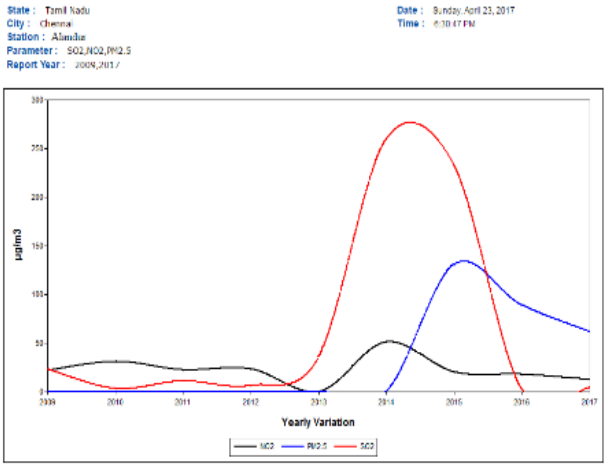


**Fig. 9:** Values of Pollutants at Alandur Chennai.

Graph captured for Visakhapatnam, Andhra-Pradesh for different pollutants from 2009-2017. As we can see values are increasing. Values which are under or less than 50 show condition is good. Other values show pollution rate is high.

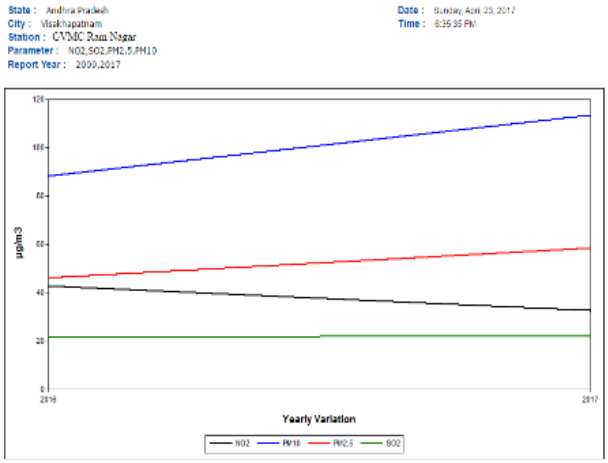


**Fig. 10:** Values of Pollutants at Vishakhapatnam (AP).

Similarly, we can find for other stations. We can compare the CPCB dashboard reports with tableau reports. As we have analyzed the CPCB portal is slower in generating the reports. It needs to be improved, which we can do by introducing tableau dashboard.

## 6. Conclusion

In this project work we have analyzed the air pollution data for Tamilnadu and Andhra-Pradesh. We have used the analytical tool and classifier to analysis the accuracy and execution time of the classifying algorithm

In this project work we analyzed the performance of classifiers and checked the accuracy of each. We have seen decision tree and random forest provide higher accuracy than other classifiers. We have seen from analysis that on Monday and Thursday pollution is high in both states and on Saturday and Sunday pollution rate is less. Tuesday and Friday pollution rate are higher than Saturday and Sunday but lesser than Monday and Thursday. Future data are predicted based on supplied data, the prediction results are of good accuracy in most of the cases. In other hand we have seen the technology comparison between python and Scala on spark platform. Scala takes less time in almost all the cases than python.

As our data was not that huge, difference is less, but if data is huge like GB or Tb it will give better results. If we connect this to live dashboard we can get pollution forecast in a faster way.

## 7. Future work

Live dashboard which shows live streaming of data and apply analysis on huge data which is flowing from different Sources. By applying different analysis and visualization techniques we can predict the future data from the previous data.

## References

[1] Carson K. Leung, Fan Jiang, Hao Zhang, and Adam G.M. Pazdor, "A Data Science Model for Big Data Analytics of Frequent Patterns" 978-1-5090-4065-0© 2016 IEEE.

[2] Mihaela Oprea, Hai-Ying Liu, "A knowledge-based approach for PM2.5 air pollution effects analysis" 978-1-4673-9910-4 ©2016 IEEE.

[3] Haripriya Ayyalasomayajula, Edgar Gabriel, Peggy Lindner, Daniel Price," Air Quality Simulations using Big Data Programming Models" 978-1-5090-2251-9/16 © 2016 IEEE.

[4] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, and Mohammad Reza Kavoosifar, "Analyzing air pollution on the urban environment" MIPRO 2016, May 30 - June 3, 2016.

[5] Navjot Kaur Walia, Parul Kalra, Deepti Mehrotra, "Prediction of Carbon Stock Available in Forest using Naive Bayes Approach" 978-1-5090-0210-8/16 © 2016 IEEE.

[6] David G. Rickerby, Andreas N. Skouloudis, "Big data for innovative air-pollution assessments in the era of verifiable regulatory decisions" 978-1-5090-0058-6/16 ©2016 European Union.

[7] Jinsong Wu, Senior Member, IEEE, Song Guo, Senior Member, IEEE, Jie Li, Senior Member, IEEE, "Big Data application in Green Challenges" 1932-8184 © 2016 IEEE.

[8] James Manyika, Michael Chui, Brad Brown, "Big data: The next frontier for innovation, competition, and productivity" Report, McKinsey Global Institute, USA, May 2011.

[9] Nancy Agrawal and Arushi Baboota, "The Importance of Including Carcinogenic Benzene in Real-Time Ambient Air Quality Data in Delhi" COMSNETS 2016 - Net Health Workshop, 978-1-4673-9622-6/16 ©2016 IEEE.

[10] Sreemoyee Roy and Abhik Mukherjee, "Information system analysis for monitoring of air quality in peri-urban Howrah" 2012 Third International Conference on Emerging Applications of Information Technology (EAIT), 978-1-4673-1827-3/12 ©2012 IEEE.

[11] Lily Bui, "Breathing Smarter: A critical look at Representation of air quality sensing data across platform and publics" 2015 IEEE.

[12] Wenjun Lv, Yu Kang, Zerui Li, Yunbo Zhao "Fusion Approach for Real-Time Mapping Street Atmospheric Pollution Concentration" 978-1-5090-1729-4/16©2016 IEEE.

[13] A. Cuzzocrea and C. K. Leung, "Computing theoretically-sound upper bounds to expected support for frequent pattern mining problems over uncertain big data," Proc. IPMU 2016, Part II, pp. 379–392.

[14] Elena Baralis, Tania Cerquitelli, Silvia Chiusano, "Analyzing air pollution on the urban environment" MIPRO 2016, May 30 - June 3, 2016, Opatija, Croatia.

[15] Abhishek Pandey, Amit Sinha, "An Analytical Approach to Check the Development of any State in India" 2016 Second International Conference on Computational Intelligence & Communication Technology, 978-1-5090-0210-8/16 © 2016 IEEE.

[16] Valerio Persico, Antonio Montieri, Antonio Pescape, "On the Network Performance of Amazon S3 Cloud-storage Service" 2016 fifth IEEE International Conference on Cloud Networking, 978-1-5090-5093-2/16 © 2016 IEEE.