Predicting Air Quality Using Machine Learning and Distributed Computing

Introduction

Due to the rapid global Industrialization and urbanization process, environmental pollution issues such as air pollution have become more and more severe. Air pollution is a mixture of solid particles and gases in the air. Major outdoor air pollutants include zone(O3), particle matter (PM), sulfur dioxide(SO2), carbon monoxide(CO) and Nitrogen oxides(NOx) [1]. Some air pollutants can trigger averse health problems, including respiratory disease, heart disease, lung cancer, brain damage, liver damage and kidney damage [2] .For instance, it has been reported that Fine particulate PM2.5 and sulfur oxide–related pollution were associated with lung cancer and cardiopulmonary mortality[3].

The problem of air pollution is much severe in California. According to the American Lung association's annual "State of the Air" report, California has the eight of the USA's 10 most-polluted cities in terms of ozone pollution and it has seven of the USA's 10 most-polluted cities both in terms of by year round particle pollution and by short-term particle pollution [4]. The air pollution issue in California is mainly caused by the regular occurrences of wildfire. According to the California Department of Forestry and Fire Protection, the average number of fires in California in five years exceeds 3500, which leads to more than 190 thousand acres has been burned and more than 600 million dollars of economical loss on average[5] . The wildfire emission can increase the concentration level of PM, CO2, NO as well as other air pollutants, resulting in a poor air quality[6]. For example, Camp Fire, the most destructive fire on record, broke out in Butte County, Northern California on November, 2018 has caused the air pollution level of Northern California ranks with the worst in the world [7]. Due to the poor air quality, several public schools in Northern California has been closed accounting for the health warning. The poor air quality can also cause a significant increase in hospital emergency room visits for asthma, respiratory problems, eye irritation, and smoke inhalation [8].

Air quality forecasting is an effective way of protecting public health by providing an early warning against harmful air pollutants. In this case, especially in California, it's of great importance for us to predict air quality to give public warning in advance and let the public sector such as schools , emergency services and healthcare to engage in pre-event planning.

The Air Quality Index (AQI) is an index for reporting daily air quality and it can be used to warn the public when air pollution is hazardous. For instance, as the AQI increases, the public is likely to experience increasing severe adverse health effects. United States Environmental Protection Agency (US EPA) calculated the AQI for five major pollutants : Particulate Matter(PM), Sulphur dioxide (SO2), Carbon monoxide(CO) , Nitrogen dioxide (NO2), ground-level Ozone (O3)[9]. Air quality can be classified into six levels according to AQI. The AQI levels of 1 to 6 indicates excellent, food, mild pollution, moderate pollution, heavy pollution, and serious pollution, respectively. According to AQI technical specification HJ633-2010(for trial implementation), Table 2 shows the AQIs and their corresponding air quality levels[9].

Table 1. AQIs and Air Quality Levels

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution posse little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for sensitive Groups	101 o 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alter : everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Since People can use the above Air Quality Information to protect themselves from unhealthy outdoor air pollution. Therefore, it's of great importance for us to predicting & forecasting air quality Index (AQI), enhancing public awareness of environment protection and improving public life quality.

In the previous, the air quality index (AQI) was calculated based on the mathematical and statistical techniques. In these techniques, a physical model is initial designed and data is coded with mathematical equations. However, these techniques usually provided limited accuracy for they are unable to predict the extreme pollution points [10]. Now, with the development of big data technology, many studies have been predicted air quality through machine learning or deep learning based model to achieve better accuracy. For example, Y. Zheng et al. [11] used both linear regression-based model and neural network based model to predict AQI over the next 48 hours in China. Yu et al. [12] applied a random forest approach for predicting air quality for urban sensing system in Sheyang, China. H. Zhao et al. [13] used an improved Artificial Neural networks (ANN) model called GA -ANN to predict AQI in Tianjin, China, in which genetic algorithm (GA) is used to select a subset of factors from the original set and the GA-selected factors are fed into ANN for modeling. L. Xiang et al . applied [14] a novel spatiotemporal deep learning (STDL)-based model to predict air quality.

However, these studies simply apply the machine learning models and seldomly utilize distributed system computing methods. As the volume of data increase, it would be computational expansive to train these machine learning and deep learning models without using distributed computing system. The distributed systems can provide high-quality aggregate performance by connecting many networked computers to compute and process each task on one or multiple machines [15]. By replicating data across multiple nodes, the system is resistant against the failure in single node[16]. Therefore, the distributed computing system enables us to process data in a fast, reliable and cost-efficient way.

Apache Spark is an open-source distributed clustering-computing framework, it extends the MapReduce model with primitives for efficient data sharing by using resilient distributed dataset (RDDs). In general, Spark has several advantages over other distributed computing systems : fast speed, ease of use, generality and runs everywhere. To be more specific , Spark has an advanced DAG execution engine which enables it to run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk. It's ease of use, because Spark can be used interactively with Scala, Java, Python and R. Spark provides Spark SQL, ML libraries for tackling machine learning problems. Spark not only can run on standalone cluster mode, it can also run on other platform like EC2, Hadoop YARN etc [17]. In this study, we would use the Apache Spark as our distributed computing system.

Though the relational databases has persistent data storage, standard model and concurrency, it fails to cope with the large scale dataset. While the NoSQL databases are more scalable and provide superior performance, for it can run well on distributed systems [18]. MongoDB is an object-oriented, simple, dynamic, and scalable NoSQL database [19]. It enables us to store and query large volume data with high availability, automatic scaling and time cost efficiency.

In this study, we focus on using Big Data technology (MongoDB, Apache Spark, AWS) and machine learning methods to model and predict California daily AQI with better computation efficiency on the basis of historical metrological data and air pollution data. By doing this, this study aims to give public warning in advance and let the public sector to engage in pre-event planning.

To achieve these, our study evaluated the Spark performance under both the different type of machine learning algorithm. We also compared the its performance with the similar processes being run on a local machine with PySpark (standalone) or with Python machine learning library

Scikit-Learn. This would help us to determine which condition is beneficial to implement distributed systems for machine learning algorithm to achieve better computation efficiency.

System Overview

System Workflow

For this study, the data science pipeline was designed around scalability, cloud resources, and distributed computing methods. In this case, we developed our pipeline by using Amazon Web Service (AWS), for AWS can provide high availability and scalability and makes its components including storage and processing engines to be compatible [20]. In general, our preprocessed data was stored in AWS Simple Storage Service(S3) bucket, then data was transferred and loaded into the MongoDB on AWS Elastic Compute Cloud (EC2). Later, our data was processed using Apache Spark on AWS Elastic MapReduce (EMR) (Fig.1).

1) Data Storage : Data scraped using APIs and preprocessed with BigQuery was stored in AWS Simple Storage Service(S3) bucket. AWS S3 is selected for it can give us highly scalable, reliable, fast cloud data storage[21]. This is because AWS S3 allows us to store and retrieve any amount of data at any time, from anywhere on the web .
2) Data Management : Since MongoDB allows us to store data of any structure and includes features such as sharding and replication, we will choose we MongoDB as our data management system. In our study, we deploy MongoDB on AWS EC2 cluster with total 10 instances. Half of instances are t2.large while the other half are t2.medium instances. AWS EC2 is a web service that provides secure, reliable, resizable capacity in the cloud [22]. It enables us to build a MongoDB cluster by automating configuration and deployment tasks [23].
3) Data Analysis: Data Processing and machine learning was performed using Apache Spark SQL and Spark ML on AWS Elastic MapReduce (EMR). AWS EMR can provide a managed Hadoop framework for processing vast amounts of data across AWS EC2 instances with an easy, fast and cost-effective setting[24].The EMR enables us to run the distributed system Apache Spark and interact with data in other AWS data stores such as AWS S3 and MongoDB built upon AWS EC2. For this research, we ran two different EMR cluster settings : one with three m4.large instances another with three m4.xlarge instances. Each setting with one instance set up as a master and the remaining two as salves. Once we connected with AWS EMR, the data was processed using the following six steps : 1) Data transfer, 2) RDD creation, 3) DataFrame creation, 4) DataFrame processing, 5) Machine Learning model training 6) Classification. The performance of EMR was tested and compared with the performance of a standard commercial laptop with 500 GB Flash Storage , 8GB RAM, and 2.9 GHZ Intel Core i5.

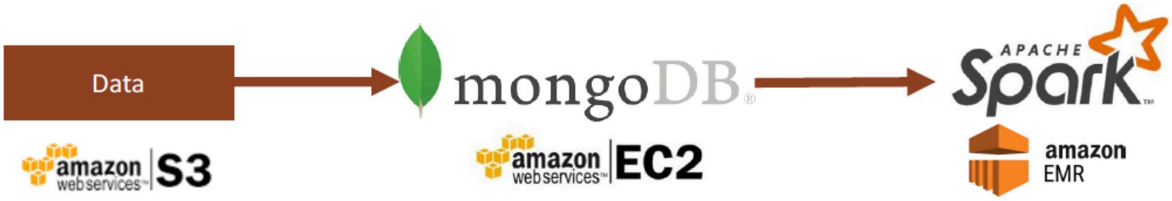

Fig 1. System work workflow

Algorithm :

In this study, we used the features (Fig. 4) to predict the daily AQI category on the city basis by using the following machine learning algorithm. Before we apply these machine learning algorithms in Spark ML, we used VectorAssembler method to merge all the new vectors and the original columns into a single vectors. It's useful combining raw features and features generated by different feature transformation into a single feature vector, in order to train Machine Learning models like logistic regression and decision tree based model. We use the following machine learning algorithm to do the classification.

a) Logistic Regression : Logistic regression is a predictive analysis used for binomial or multinomial classification problem [26]. Logistic Regression minimizes a linear combination of our input features through gradient descent. But unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete class [27].

b) Random Forest Classification : Random Forest Classification is an ensemble method that fits multiple decision tree classifiers on various bootstrap samples (re-sample the data many times with replacement and re-estimate the model) of the dataset [28]. The output class is the mode of the classes of the individual trees. Due to the instability of single regression tree, random forest classification can improve the predictive accuracy as well as control the overfitting to their training problem.

Experiment Output :

A. Data :

The data that we use is the AQS Data Mart[30] , which contains the historical air quality information across the state. It also includes the associated aggregate values calculated by EPA

(8-hour, daily, annual, etc.) The factors include air pollutants that would affect AQI, such as CO, NO_2, O_3 , pm10, pm 2.5 firm, pm 2.5 Non-firm, pm 2.5 speciation and SO_2. It also has other factors like temperature, wind, and pressure. In our project, we only use the daily data in California to run the classification analysis.

We utilized the BigQuery Python client library and BigQuery API to query the data. The query language is SQL. According to the city name and date, the tables of different features are joined. The full table was obtained after the SQL query. Since our goal is to predict the AQI level on the city basis, the data was then grouped by the same city to get the mean value of each air pollutants, temperature, wind, pressure as well as the max AQI value for each city.

Our preprocessed data includes 600 thousand observations of California air quality information collected between the year 2008 and the year 2017 with the size of around 1 GB. This large dataset would enable us to evaluate the computation efficiency of distributed computing system.

As our objective was to benchmark performance rather than achieve marginal gains in prediction accuracy, we conducted minimal feature engineering. We basically used the previous day concentration of air pollutants with their associated AQI and other influential factors like temperature, wind to predict the classification of the AQI. The more detailed features depicted in the Fig .2

```
root
 |-- aqi_co: double (nullable = true)
 |-- aqi_no2: double (nullable = true)
 |-- aqi_o3: double (nullable = true)
 |-- aqi_pm10: double (nullable = true)
 |-- aqi_pm25_frm: double (nullable = true)
 |-- aqi_pm25_nonfrm: double (nullable = true)
 |-- aqi_so2: double (nullable = true)
 |-- arithmetic_mean_co: double (nullable = true)
 |-- arithmetic_mean_no2: double (nullable = true)
 |-- arithmetic_mean_o3: double (nullable = true)
 |-- arithmetic_mean_pm10: double (nullable = true)
 |-- arithmetic_mean_pm25_frm: double (nullable = true)
 |-- arithmetic_mean_pm25_nonfrm: double (nullable = true)
 |-- arithmetic_mean_pm25_speciation: double (nullable = true)
 |-- arithmetic_mean_pressure: double (nullable = true)
 |-- arithmetic_mean_so2: double (nullable = true)
 |-- arithmetic_mean_temp: double (nullable = true)
 |-- arithmetic_mean_wind: double (nullable = true)
 |-- city_name: string (nullable = true)
 |-- county_code: integer (nullable = true)
 |-- day: integer (nullable = true)
 |-- dow: integer (nullable = true)
 |-- first_max_value_co: double (nullable = true)
 |-- first_max_value_no2: double (nullable = true)
 |-- first_max_value_o3: double (nullable = true)
 |-- first_max_value_pm10: double (nullable = true)
 |-- first_max_value_pm25_frm: double (nullable = true)
 |-- first_max_value_pm25_nonfrm: double (nullable = true)
 |-- first_max_value_pm25_speciation: double (nullable = true)
 |-- first_max_value_pressure: double (nullable = true)
 |-- first_max_value_so2: double (nullable = true)
 |-- first_max_value_temp: double (nullable = true)
 |-- first_max_value_wind: double (nullable = true)
 |-- label: string (nullable = true)
 |-- latitude: double (nullable = true)
 |-- longitude: double (nullable = true)
 |-- max_aqi: double (nullable = true)
 |-- max_aqi_before_yesterday: double (nullable = true)
 |-- max_aqi_yesterday: double (nullable = true)
 |-- month: integer (nullable = true)
 |-- observation_count_co: double (nullable = true)
 |-- observation_count_no2: double (nullable = true)
 |-- observation_count_o3: double (nullable = true)
 |-- observation_count_pm10: double (nullable = true)
 |-- observation_count_pm25_frm: double (nullable = true)
 |-- observation_count_pm25_nonfrm: double (nullable = true)
 |-- observation_count_pm25_speciation: double (nullable = true)
```

Fig 2. Features

Results :

To evaluate the performance of our models, we not only evaluate their fitting time, we also compare their accuracy and F1 score. Accuracy is the fraction that our model predict correctly. Basically, in Figure 3, the Random Forest performs better than the Logistic Regression according to the accuracy. In this case, the number of trees doesn't affect the accuracy and F1 score.

As the parameters changes, the supervised algorithms could become more and more complex, thus taking more time for the machine to train the models. As the complexity increases, the bigger challenge would be faced by the machines for the memory limits and time availability.

For the two algorithms, the Logistic Regression could deal with large size of data relatively easily. Overall, this algorithm takes shorter time and requires less memory to train the model. However, as the number of estimators changes, the Random Forest algorithm could become very complex. Basically, when the number of estimators grows, it would take longer and longer time to train the model. There are three choices of tree numbers and can be seen in figure 4: 10, 100 and 500. It is obvious to observe that when in the same context, as the number of trees increases, the total time becomes longer.

The algorithms are applied in four contexts(PySpark(standalone), EMR Cluster 8G, EMR Cluster 16G, Scikit-Learn), the times are recorded and shows in the figure 4 and figure 5. The algorithms in scikit-learn take the longest time. The unit of measurement is second(s). The distributed computing based algorithms take really shorter time which is measured in millisecond(ms). After employing two types of EMR clusters, comparing to the PySpark standalone situation, it shows the time is reduced significantly when more memory and cores are available.

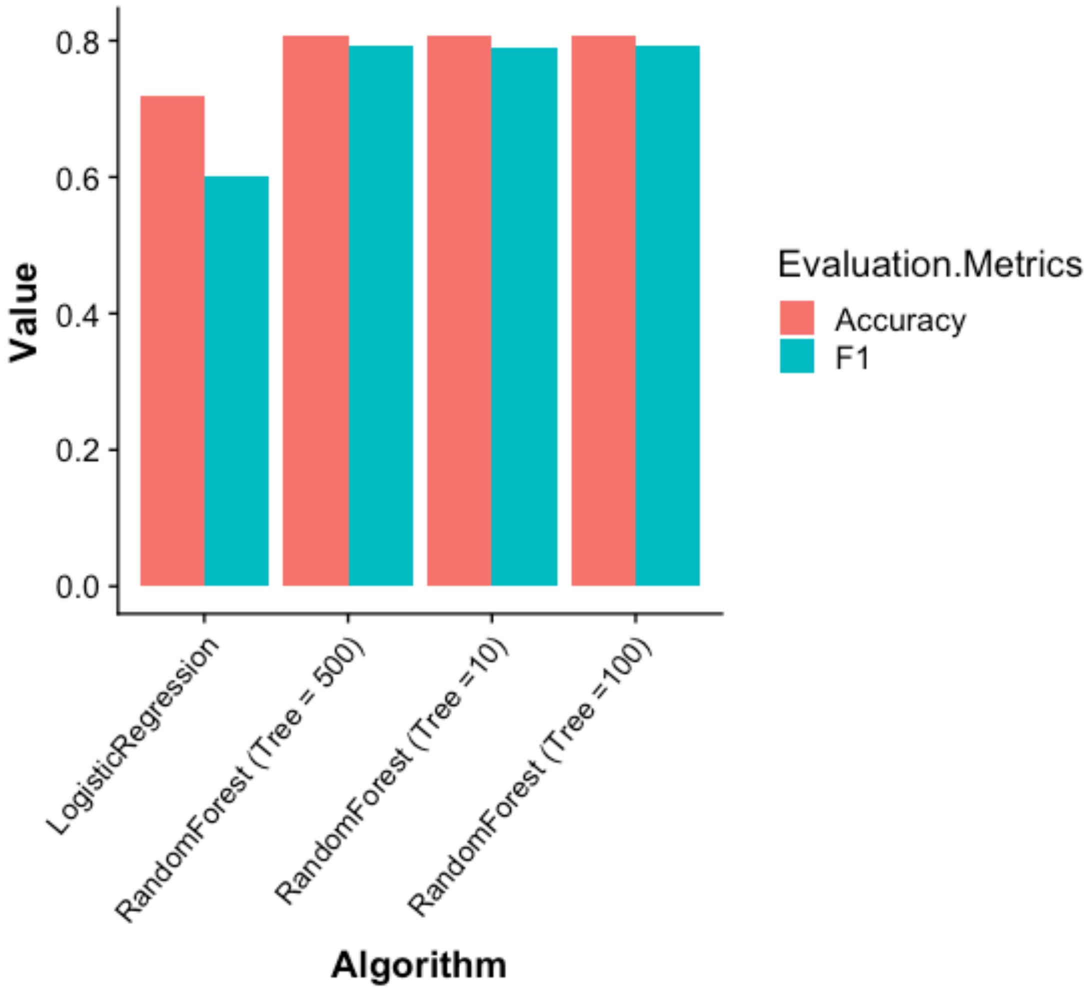

Fig. 3 Models Evaluation on EMR Cluster with 8GB RAM.

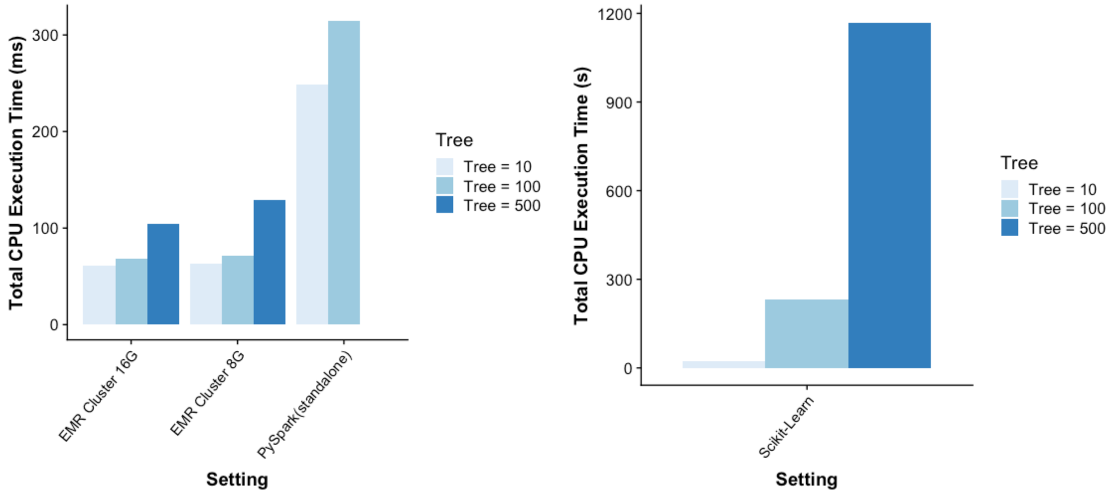

Fig 4. Comparison of Random Forest Regressor fit times for Cluster vs PySpark (standalone) vs Scikit-Learn

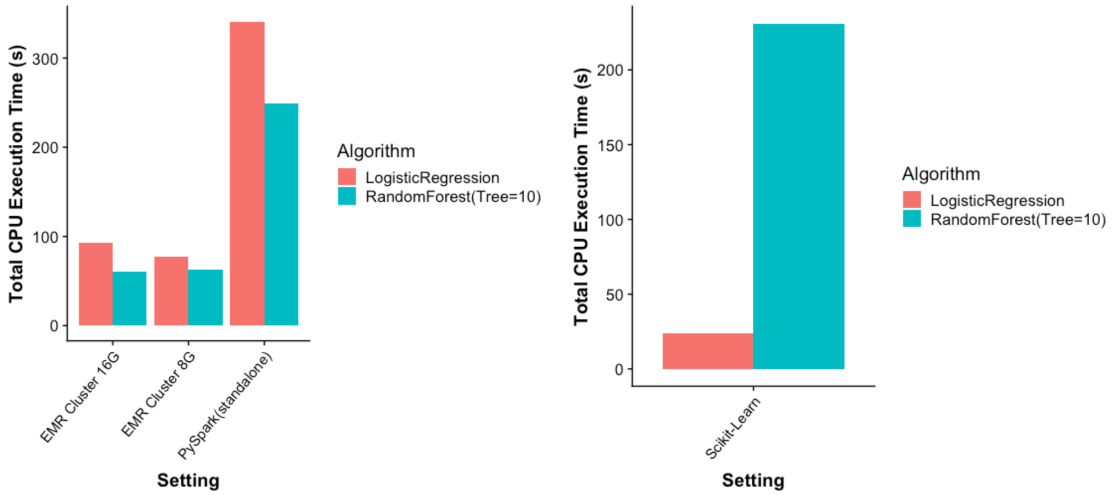

Fig 5. Comparison of Model Training Time for Cluster vs PySpark (standalone) vs . Scikit-Learn

References :

1. Curtis, Luke, et al. "Adverse health effects of outdoor air pollutants." Environment international 32.6 (2006): 815-830.
2. Samet, Jonathan M., et al. "The national morbidity, mortality, and air pollution study." Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst 94.pt 2 (2000): 5-79.
3. Pope III, C. Arden, et al. "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution." Jama 287.9 (2002): 1132-1141.
4. American Lung Association, ‘Most Polluted Cities’ , 2018. [Online]. Available : https://www.lung.org/our-initiatives/healthy-air/sota/city-rankings/most-polluted-cities.html
5. California Department of Forestry and Fire Protection, ‘ CAL FIRE Jurisdiction Fires, Acres, Dollar Damage, and Structures Destroyed (1933-2016)’, 2018. [Online]. Available : http://cdfdata.fire.ca.gov/pub/cdf/images/incidentstatsevents_270.pdf
6. National Research Council: Committee on Air Quality Management in the United States, Board on Environmental Studies and Toxicology, Board on Atmospheric Sciences and Climate, Division on Earth and Life Studies (2004). Air Quality Management in the United States. National Academies Press. ISBN 0-309-08932-8
7. San Francisco Chronicle, ‘ Northern California air quality rated the worst in the world, conditions ‘hazardous’, 2018. [Online]. Available : https://www.sfchronicle.com/california-wildfires/article/Smoke-still-plagues-Bay-Area-skies-a-week-after-13394932.php

8. Delfino, Ralph J., et al. "The relationship of respiratory and cardiovascular hospital admissions to the southern California wildfires of 2003." Occupational and environmental medicine66.3 (2009): 189-197
9. US EPA, "Air Quality Index(AQI) – A Guide to Air Quality and Your Health", 2011. [Online]. Available : https://www.airnow.gov/index.cfm?action=aqibasics.aqi
10. V. M. Niharika and P. S. Rao, "A survey on air quality forecasting techniques," International Journal of Computer Science and Information Technologies, vol. 5, no. 1, pp.103-107, 2014
11. Zheng, Yu, et al. "Forecasting fine-grained air quality based on big data." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.
12. Zhao, Hong, et al. "A GA-ANN model for air quality predicting." Computer Symposium (ICS), 2010 International. IEEE, 2010.
13. Yu, Ruiyun, et al. "Raq–a random forest approach for predicting air quality in urban sensing systems." Sensors 16.1 (2016): 86.
14. Li, Xiang, et al. "Deep learning architecture for air quality predictions." Environmental Science and Pollution Research23.22 (2016): 22408-22417.
15. Coulouris, George; Jean Dollimore; Tim Kindberg; Gordon Blair (2011). Distributed Systems: Concepts and Design (5th Edition). Boston: Addison-Wesley
16. Howard, Alexander, et al. "Distributed Data Analytics Framework for Smart Transportation." 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, 2018.
17. Apache Spark, "Apache spark: Lightning-fast cluster computing," 2018. [Online]. Available: http://spark.apache.org
18. Chodorow, Kristina. MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. " O'Reilly Media, Inc.", 2013
19. MongoDB. (2018) Mongodb for giant ideas. [Online]. Available:https://www.mongodb.com/
20. https://en.wikipedia.org/wiki/Amazon_Web_Services
21. https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html
22. https://aws.amazon.com/ec2/
23. https://docs.aws.amazon.com/quickstart/latest/mongodb/overview.html
24. https://aws.amazon.com/emr/
25. Air Quality Index Wikipedia. [Online]. Available : https://en.wikipedia.org/wiki/Air_quality_index#Computing_the_AQI
26. J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, Applied linear statistical models. Irwin Chicago, 1996, vol. 4.
27. J. S. Cramer, "The origins and development of the logit model," Logit models from economics and other fields, vol. 2003, pp. 1–19, 2003.
28. Barandiaran, Iñigo. "The random subspace method for constructing decision forests." *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998).
29. AQS Data Mart Website. [Online]. Available : https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html