 CMC, vol.59, no.3, pp.697-711, 2019

Development of Cloud Based Air Pollution Information System Using Visualization

SangWook Han[1], JungYeon Seo[1], Dae-Young Kim[2], SeokHoon Kim[3] and HwaMin Lee[3, *]

Abstract: Air pollution caused by fine dust is a big problem all over the world and fine dust has a fatal impact on human health. But there are too few fine dust measuring stations and the installation cost of fine dust measuring station is very expensive. In this paper, we propose Cloud-based air pollution information system using R. To measure fine dust, we have developed an inexpensive measuring device and studied the technique to accurately measure the concentration of fine dust at the user's location. And we have developed the smartphone application to provide air pollution information. In our system, we provide collected data based analytical results through effective data modeling. Our system provides information on fine dust value and action tips through the air pollution information application. And it supports visualization on the map using the statistical program R. The user can check the fine dust statistics map and cope with fine dust accordingly.

Keywords: Air pollution, visualization, R, big data, cloud clusters.

1 Introduction

Since the end of 2000, smartphones have proliferated and have enriched people's lives. With the proliferation of smartphones, new smart devices such as tablet PCs, smart TVs, smart refrigerators, and smart air conditioners have emerged, expanding from personal to business and home [Kim, Moon and Park (2017)]. With the spread of smart devices and the rise of SNS, massive data is being produced. It reached the age of no longer living without data. Big Data can be said to change the world.

Big Data represents a paradigm shift that involves not only the management activities of private companies but also the public sector, including the government. It is also used to solve some national social issues such as disaster, social safety, welfare, and medical care. In other words, through an objective, scientific approach based on data, it is possible to cope with the problem of national society and the change of future society [Yoon (2013)].

Recently, cluster environment of cloud computing market is changing from a general cluster environment to virtual cluster environment. These changes in the cluster environment affect the performance of large capacity distributed processing. Distributed

[1] Deptartment of Computer Science and Engineering, Soonchungyang University, Asan, 31538, Korea.

[2] School of Information Technology Engineering, Daegu Catholic University, Gyeongsan, 38430, Korea.

[3] Department of Computer Software and Engineering, Soonchungyang University, Asan, 31538, Korea.

[*] Corresponding Author: HwaMin Lee. Email: leehm@sch.ac.kr.

processing of large amounts of data can be expected to have advantages such as resource management efficiency, reliability, equipment purchases cost reduction, and power cost reduction. Many IT companies outside Korea are investing heavily in research and services competitively [Li and Wang (2017)].

In this paper, we develop air pollution information system through fine dust big data which is emerging as a social problem in Korea. Interest and risk for fine dust are increasing. However, due to the lack of a fine dust measuring station as shown in Tab. 1, the user is hardly provided with information on the fine dust value. In Korea, for example, there are only four stations (Chuncheon, Wonju, Donghae, Samcheok) that provide fine dust value among 18 cities in Gangwon Province. More than half of the measuring station (7 places) all over Gangwon Province locate in Chuncheon (2 places) and Wonju (2 places). The following Tab. 1 shows the area of the jurisdiction per site of the fine dust measuring station [Franceschi, Cobo and Figueredo (2018)]. Since the fine dust measuring station is mainly located in the metropolitan area, it is difficult to know the accurate fine dust information in the suburbs. Nationwide, fine dust measuring stations are fewer than kindergartens and elementary schools. This is a big social problem.

Table 1: Fine dust measurement area by region (unit: km^2)

City	City area	Number of measuring stations	Area per one
Seoul	605.3	39	15.5
Incheon	575.4	18	32.0
Gyeonggi	3357.5	80	42.0
Busan	940.8	21	44.8
Daejeon	495.5	10	49.6
Ulsan	755.6	15	50.4
Gwangju	480.1	9	53.3
Jeonbuk	885.5	15	59.0
Daegu	798.0	13	61.4
Chungbuk	724.5	11	65.9
Sejong	141.0	2	70.5
Gyeongnam	1896.3	20	94.8
Jeonnam	1729.0	16	108.1
Chungnam	903.4	8	112.9
Gyeongbuk	1850.2	14	132.2
Gangwon	1022.6	7	146.1
Jeju	453.2	3	151.1
Total	17613.7	301	58.5

In this paper, we use the low-cost fine dust measurement sensor to collect accurate fine dust data and inform the user through fine dust information application. Also, the database is efficiently designed by modeling the collected data. Our system statistically analyzes fine dust Big Data using R program. Visualization improves readability and provides users with fine dust information more efficiently. Also, the user can know the fine dust value of the current position. When the user confirms the fine dust information application, they can wear a mask or prevent outdoor activities. In the future, we can learn the pattern by learning the fine dust data and predict the path. This monitoring system can be applied to environmental or health problems.

2 Related work

2.1 Data modeling

The process of creating from the real world to the database proceeds according to time as shown in Fig. 1. The conceptual data model, the logical data model, and the physical data model are organized according to the level of abstraction [Chung and Nah (2018); Ratliff, Balise, Veerayahu et al. (2016)].

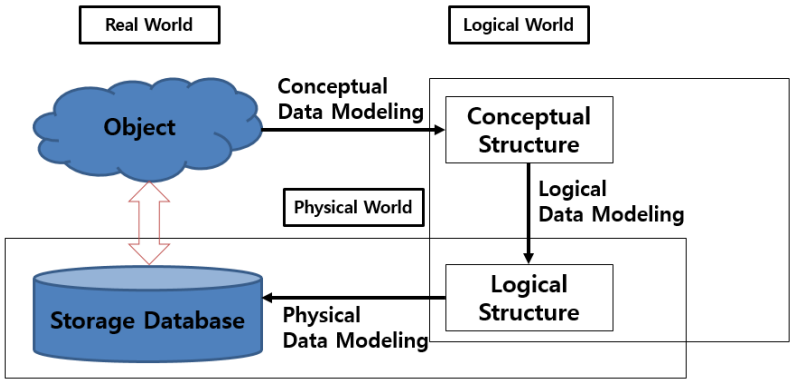

Figure 1: A model between the real world and the database

2.1.1 Conceptual data modeling

Conceptual data modeling begins with finding and analyzing the data requirements of an organization or user. This process includes determining what data is important and what data should be maintained. The main activity at this stage is to discover the relationships between the core entities and their relationships and to create entity-relationship diagrams to represent them. Entity-relationship diagrams are used to indicate what data is important to the organization and to various database users. Formulating an organization's data needs through the conceptual data model supports two important functions. First, the conceptual data model supports the discovery of data requirements by users and system developers. The conceptual data model is abstract. Therefore, the model makes it easy to structure the upper problems and forms the basis for users and developers to discuss system functions. Second, the conceptual data model is useful for

understanding how the current system should be transformed. In general, very simple isolated systems are more easily represented and explained through abstract modeling

2.1.2 Logical data modeling

Logical data modeling is a key part of the data modeling process. It is a technique or process that expresses the logical structure and rules of information. The logical data model resulting from logical data modeling is to identify and record facts that exist in business data, independent of who accesses the data and is computerized. Another important activity performed at this stage is normalization. Normalization is a process related to the process of improving the relationship of the relational data model to more structured through various types of tests. It is the process of designing a better table schema by minimizing duplication of data and eliminating the anomaly caused by data update. In normalization, the table is decomposed into a desirable form to solve the problems of the table. Through normalization, satisfy the constraints given at each step to create some normal forms in order [Jang (2007)]. The normalization process is divided into 1, 2, 3, 4, 5, and BCNF normalization form. Fig. 2 shows the relationship between normalization.

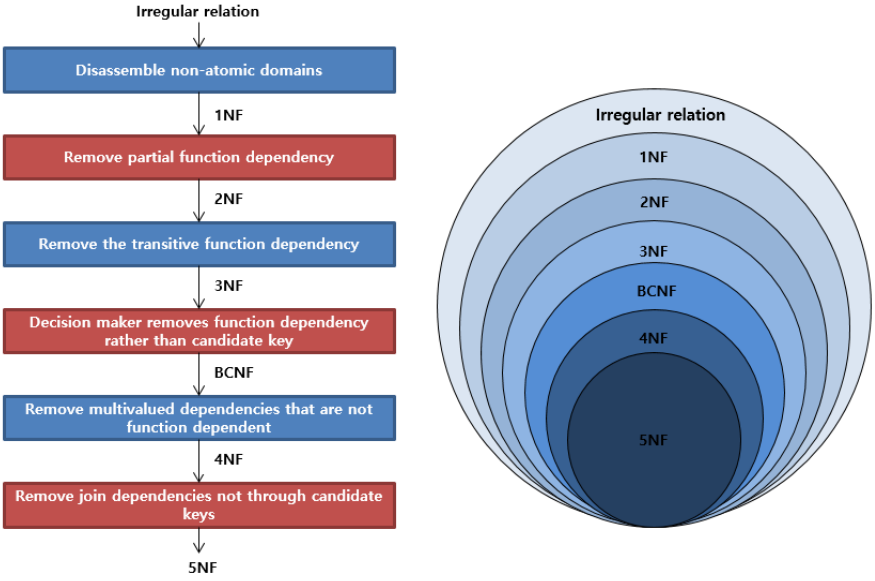

Figure 2: Relationship between normalization

2.1.3 Physical data modeling

The third phase of the database design process, physical data modeling, deals with how the logical data model will be represented on the computer hardware as a data store. The definition of how data will be physically stored on the computer is called the physical schema. What is determined at this stage is the physical storage structure represented by tables, columns, etc., the storage device to be used, and the approach to be used to extract the data. In a hierarchical database management system environment, the database

administrator must spend more time designing and implementing the physical schema. Tab. 2 describes the data model.

Table 2: Explanation of data modeling

Data modeling	Explanation
Conceptual Data Modeling	-High level of abstraction, work-oriented and comprehensive modeling. -Mostly used for EA establishment.
Logical Data Modeling	-Accurately express key, attribute, relationship, etc. for the task to build system. -Highly reusable.
Physical Data Modeling	-Stored to be portable to the database. -Designed considering physical characteristics.

2.2 R

R programming language is a programming language for statistic computation and graphics and is currently popular with big data. Based on collected data, data related to the sharing economy was examined by word-cloud. Through analyzing the increase in the amount of data from year to year, this paper will examine the level of interest in the sharing economy [Kim, Yun, Jung et al. (2018)]. Tab. 3 shows a comparison of statistical programs.

SAS and SPSS are capable of large-scale processing and provide formal support. The big disadvantage is expensive and slow. However, R is a public statistical package and free. There is no formal support, but the community is active among users so that anyone can easily access and share information. In the field of visualization, R has a wide range of available modules, making it easy to customize and optimize. R has been widely commercialized because of its advantages [Bollhuis and Kromme (2016)].

2.3 Related works

Finogeev et al. [Finogeev, Parygin and Finogeev (2017)] studies intensive approaches to distributed sensor data processing. In this paper, they propose a method to process for monitoring objects spatially distributed through GRID computing model. The GRID computing model is better suited for heterogeneous information from sensor data, integrated indicators and other sources (e.g., weather stations, security and fire alarm systems, video surveillance systems, etc.). This method reduces the load on the server cluster and reduces the amount of traffic on the sensor network.

Table 3: Statistical program comparison

	Advantages	Disadvantages
SAS	1. High adoption rate in major industries 2. Flow-based interface with drag and drop 3. Official support 4. Handling large dataset	1. Relatively high cost 2. Slow adapting to new techniques 3. Different programs for visualization or Data Mining 4. Writing code
SPSS	1. Used a lot of universities 2. Good user interface 3. Writing code made easy 4. Official support	1. Relatively high cost 2. Different licenses 3. Syntax limited 4. Slow in handling a large dataset 5. Slow adapting to new techniques
R	1. Free 2. Big community 3. Early adopter in explanatory 4. Easy to connect to data source	1. Steep learning curve 2. No official support 3. No user interface

Ahn et al. [Ahn, Jung and Park (2014)] proposes a smart air quality monitoring system using sensing data. Air quality monitoring is carried out using asbestos, which is often used in construction sites. In this paper, optical properties of chrysotile, mostly observed asbestos in the indoor atmosphere, was analyzed by spectral sensor and chromate-based on the distinctive change according to use of reagents such as water and refractive index liquid. It is research that can monitor the air quality at an economical cost using a wireless sensor network. However, since only asbestos can be monitored, it is less used in real life.

Derdour et al. [Derdour, Kechar and Fayçal-Khelfi (2016)] proposes a method for efficient data collection, which is a major task in a wireless sensor network. Mobile sensors have difficulty in collecting when they are inaccessible or at a distance. We propose a collection approach based on relay nodes and a probabilistic based mobility model to collect data efficiently. This data collection technique can be applied to critical applications such as natural disasters, environmental monitoring, and health monitoring.

Li et al. [Li, Kim and Shin (2016)] proposes a Geohash-based spatial index method for monitoring biometric data at any time using various types of wearable or embedded sensors. The proposed spatial index method can efficiently process location-based queries for medical signal monitoring. The real-time location data of the patient is constructed, and the record data is managed using the B-tree based local tree. This results in efficient location-based monitoring systems.

Li et al. [Li, Li, Chen et al. (2018)] propose opportunistic data offloading solutions to mitigate traffic on mobile networks. The explosive increase in the number of wireless data is becoming increasingly burdensome in wireless networks today. A critical approach was used to select some important users as seeds and to discover the ties of the surrounding data. Based on Gaussian graphical modeling, research was conducted to find out important factors to reduce data overload.

3 System architecture

In this paper, we propose a system structure that can measure the concentration of fine dust at user location using low-cost sensor and provide various services using measured data. Fig. 3 shows the concreted cloud system needed to analyze the large-scale sensing data measured in real-time and provide the analyzed result as a service.

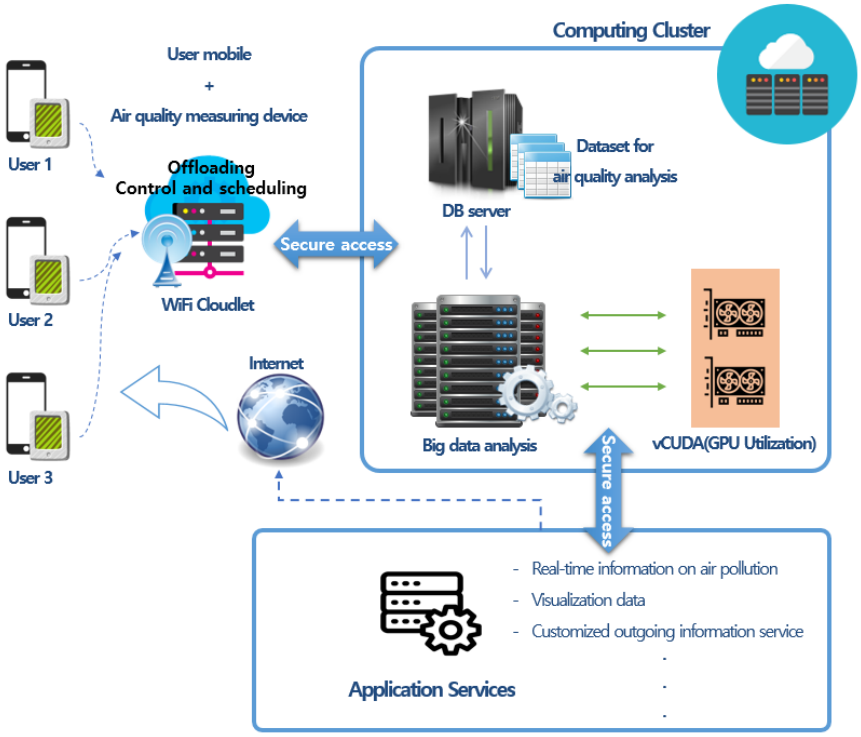

Figure 3: Cloud computing environment for visualization services

We have constructed a cloud cluster environment to collect and analyze about 200,000 pieces of fine dust data and derive the visualization data using R. We provide the calibration, inference, and API information from the cloud server when RAW data, smartphone GPS and IMU information from our device are delivered to the backend system.

3.1 Framework

The inaccurate data measured by low-cost sensors can be converted into accurate data by using the deep learning, and the contamination level in the absence of the device can be deduced by using the measured minute dust concentration of each region. Fig. 4 shows the framework proposed in this paper. In this paper, the measured data is calibrated by using the developed device, and the map is derived by using the calibrated data. Because the concentration of fine dust is greatly affected by weather factors such as temperature, humidity, pressure, etc., it is necessary to mix various data. It collects POIs, public data, personal data, meteorology data, corrects the accuracy using the artificial neural network, and provides the corrected data as a service. Existing air quality measurement products are very expensive, but also have high error rates. Using the framework developed in this study, many people can use the inexpensive sensor, and the prediction range of fine dust information becomes wider.

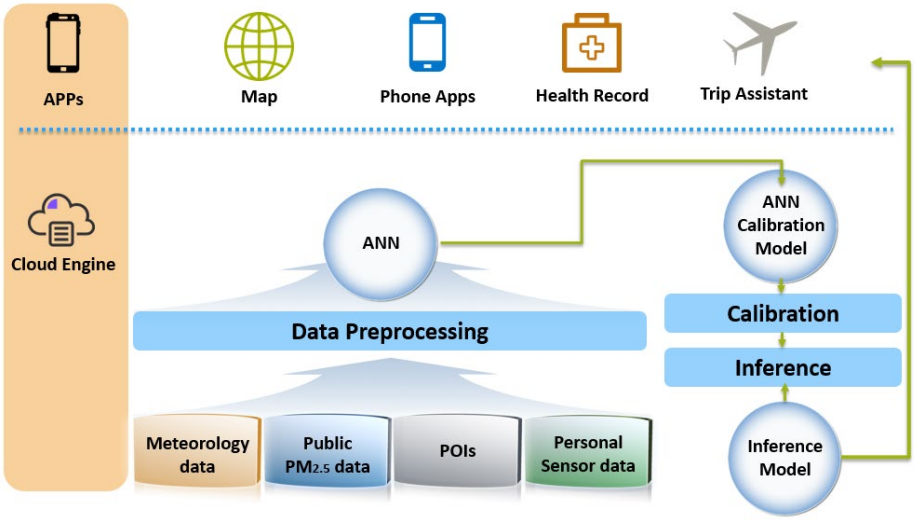

Figure 4: The proposed framework

3.2 Data modeling

Data modeling consists of three phases: conceptual modeling, logical modeling, and physical modeling. Data modeling enables more efficient R analysis.

Fig. 5 shows the logical data modeling. There are latitude and longitude for collecting location information, time information for checking the fine dust value with time, and actual fine dust and ultrafine dust value.

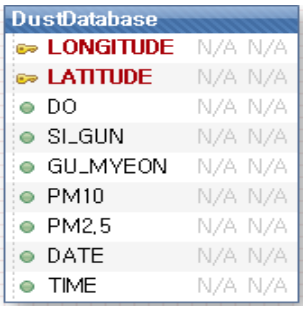

Figure 5: Logical data modeling

Fig. 6 is physical data modeling. This is set for fine dust measuring device. This data model is analyzed to inform the user of the fine dust value. It shows the values of fine dust and ultrafine dust collected by time and region.

Filter

	DATE	TIME	DO	SI_GUN	GU_MYEON	PM10	PM2.5	LONGITUDE	LATITUDE
1	20160101	100		Seoul	GangNam	100.00000	67	127.0373	37.5173
2	20160101	200		Seoul	GangNam	103.00000	60	127.0373	37.5173
3	20160101	300		Seoul	GangNam	89.00000	59	127.0373	37.5173
4	20160101	400		Seoul	GangNam	91.00000	52	127.0373	37.5173
5	20160101	500		Seoul	GangNam	87.00000	53	127.0373	37.5173
6	20160101	600		Seoul	GangNam	83.00000	46	127.0373	37.5173
7	20160101	700		Seoul	GangNam	93.00000	50	127.0373	37.5173
8	20160101	800		Seoul	GangNam	93.00000	47	127.0373	37.5173
9	20160101	900		Seoul	GangNam	89.00000	45	127.0373	37.5173
10	20160101	1000		Seoul	GangNam	89.00000	49	127.0373	37.5173
11	20160101	1100		Seoul	GangNam	86.00000	51	127.0373	37.5173
12	20160101	1200		Seoul	GangNam	93.00000	53	127.0373	37.5173
13	20160101	1300		Seoul	GangNam	97.00000	71	127.0373	37.5173
14	20160101	1400		Seoul	GangNam	89.00000	64	127.0373	37.5173
15	20160101	1500		Seoul	GangNam	73.00000	52	127.0373	37.5173
16	20160101	1600		Seoul	GangNam	44.00000	31	127.0373	37.5173

Figure 6: Physical data modeling

4 Implementation

4.1 Measuring device & experiment environment

First, a fine dust measuring device for measuring fine dust data is required. Fig. 7 shows the fine dust measuring device developed in this paper. It is convenient because smaller in size than existing fine dust measuring station and can be supplied with power by making it into an auxiliary battery case.

Figure 7: Our fine dust measuring device

The fine dust level measured in real time in our fine dust measuring device is very sensitive to wind or position. To obtain accurate fine dust data using the low-cost sensor, we constructed the experimental environment as shown in Fig. 8. Through the experiment as shown in Fig. 8, we investigated how the fine dust measured according to the wind, temperature, and sensor location changes. Since there is no way to define the relationship between measured values and accurate air quality values of a sensor with low accuracy, it is necessary to define the relationship between nonlinear data based on an artificial neural network. In order to measure accurate air quality data, it is possible to reduce the error range of low-level sensor by learning meteorology, LAW data, and public data with an artificial network.

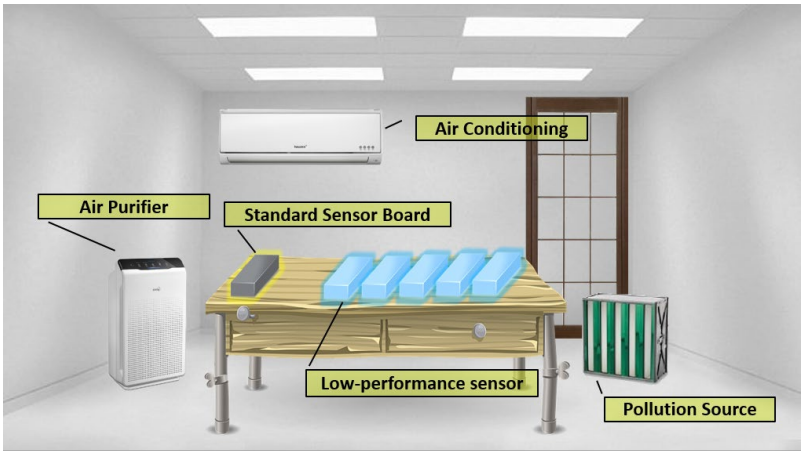

Figure 8: Experimental environments for fine dust measurement

In the environment where the fine dust concentration can freely change, the calibration work was performed so that the low-cost sensor can measure the accurate data by using the reference board which can measure the accurate data. Fig. 9 shows the process of calibrating a low-cost sensor. Using the experimental results in Fig 8, we proposed a calibration process as shown in Fig. 9 to obtain more accurate fine dust measurement values. In an artificial neural network, public PM2.5 data is given as an input value and has Ground truth value given by the standard sensor board as output.

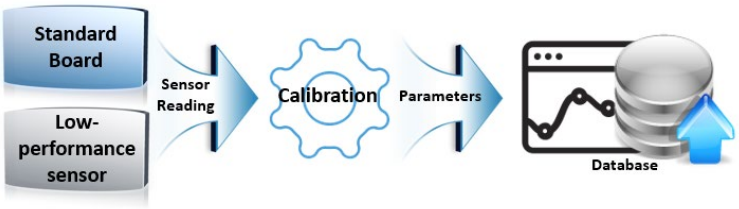

Figure 9: Calibration process

4.2 Fine dust information application

Existing systems have three types of good, normal, and bad, which simply indicates the degree of fine dust. However, this information does not tell us what action the user should take. Thus, we provide a user with a fine dust information application developed by the present inventors. Through the application, weather information such as temperature, humidity, fine dust, wind speed, ozone, ultraviolet ray, and precipitation can be checked. It defines the behaviors and criteria that users should take when going out in dozens of conditions by analyzing according to the fine dust value. Fig. 10 shows the UI of fine dust information application.

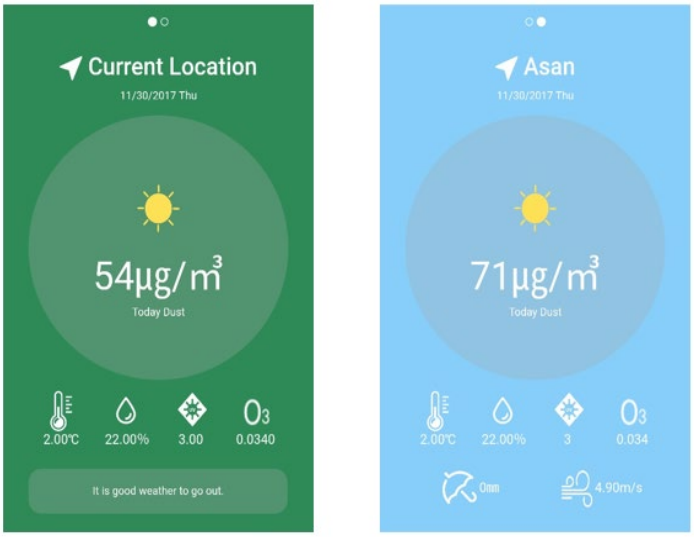

Figure 10: Our air pollution information application

4.3 Implementation of map service using R

There are about 200,000 fine dust data to be visualized. This is the result of collecting fine dust data for one year in 2016 in Seoul area. In the fine dust statistics map shown using R, the X-axis represents the longitude, and the Y-axis represents the latitude. The collected fine dust and ultrafine dust are divided into area units, and the average is performed per month to visualize. We implemented the map service using the google

map API. The google map API is a location-based service that automatically displays the local language (Korean) and English.

Fig. 11 shows April, which is the highest month of the average fine dust value (PM 10). This is the month with the highest fine dust values in a year. In April, the average of fine dust value is above 90. While the lowest month is July. The fine dust value in July is 45. April is twice as high as in July. In Korea, when the value of fine dust is over 80, air quality is bad, and it is avoided. This is a very bad result when compared with April.

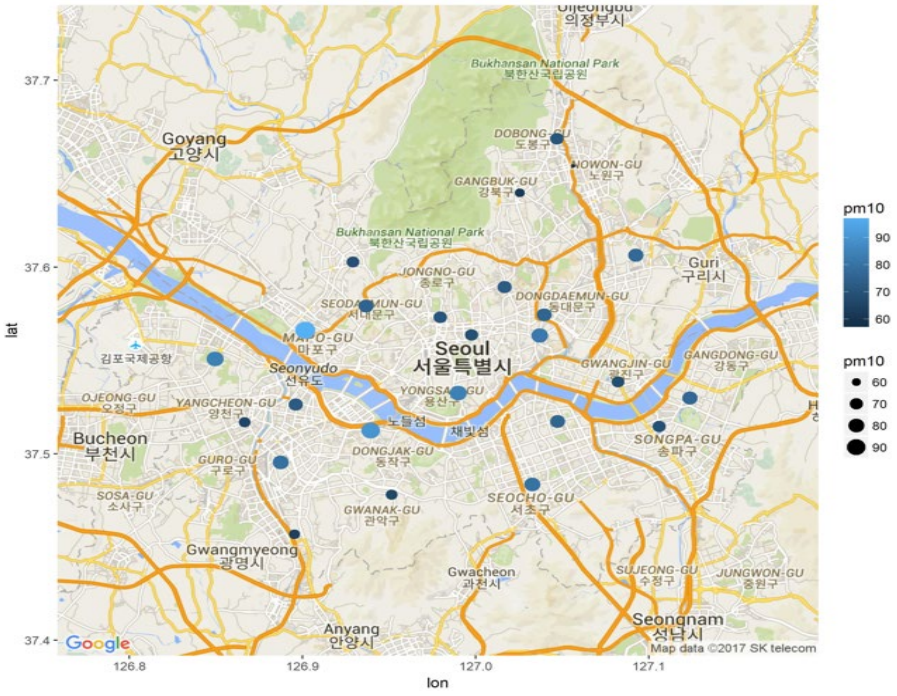

Figure 11: Month with the highest fine dust (PM 10) index: April

Fig. 12 shows March, which is the highest month of the average ultrafine dust value (PM 2.5). This is the month with the highest ultrafine dust values in a year. In March, the average of ultrafine dust value is above 39. On the other hand, the lowest month is July. The ultrafine dust value in July is 25. March is 1.56 times higher than in July. In Korea, when the value of ultrafine dust is over 40, air quality is bad, and it is avoided. This is a very bad result when compared with March. Experimental results show that the highest fine dust month is April, and the highest ultrafine dust month is March. It concludes that it is related to the yellow dust, which is the main cause of spring air pollution. The lowest month for fine dust and ultrafine dust is July. It is concluded that this was related to the rainy season in summer.

In this paper, statistical analysis is performed using R to analyze fine dust Big Data, and it is shown on the map by improving readability. Experimental results showed different indices for each region and each day. The analysis also shows that the amount of fine dust changes is high and low.

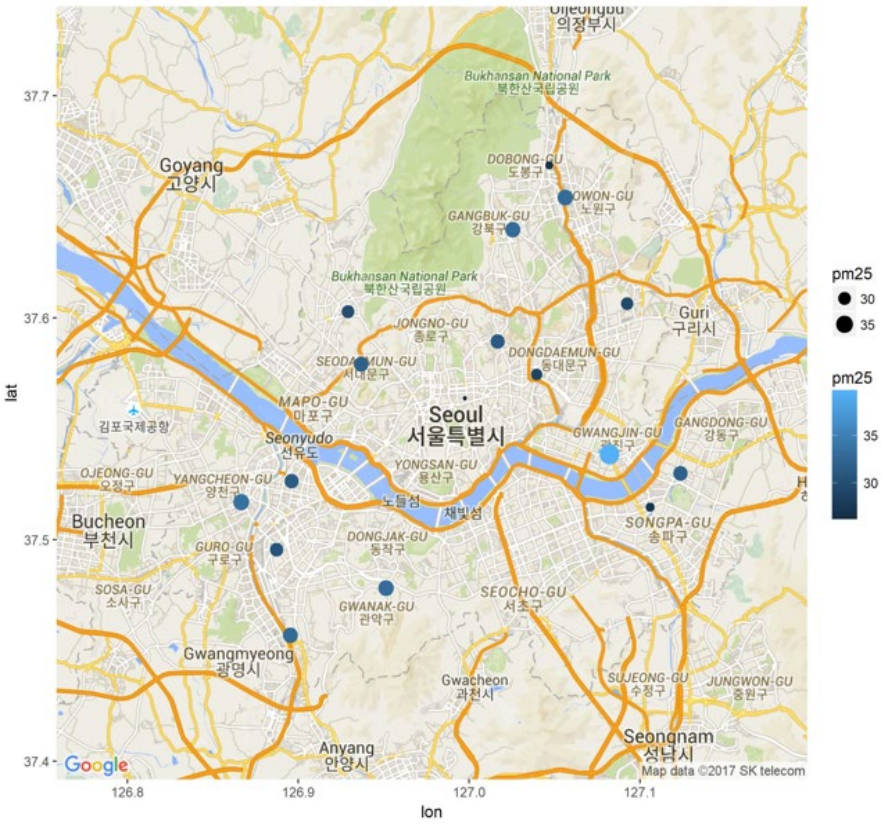

Figure 12: Month with the highest ultrafine dust (PM 2.5) index: March

5 Conclusion and future work

We have developed a system that provides air pollution information such as fine dust and ultrafine dust. With our system, users can know precisely the fine dust figure based on user location. It can be a substitute for a high-cost, low-supply fine dust gauge. Also, the cluster computing system can be used to distribute a large amount of fine dust data, thereby achieving the advantages of resource management efficiency, reliability, reduction of equipment cost, and power cost reduction. Through data modeling, it is possible to design efficiently large-scale fine dust database. Unlike existing systems, fine dust can easily be prevented by using fine dust information application. R can be used to obtain more efficient analysis results. With visualization processing, users can visually recognize the risk of fine dust again. Readability is improved, and anyone can easily check the analysis result.

In future work, we will develop the prediction system of air pollution to collect and analyze a larger amount of data from the fine dust measurement device. The collected data is learned by using deep learning. Based on the learned data, we will propose a prediction algorithm of the movement path pattern of the fine dust. It can provide the user with the expected movement path of the fine dust through visualization. Based on these

predictions, users will be able to prevent air pollution. In addition to air pollution, can be used for other environmental pollution and health monitoring systems.

Acknowledgment: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1A2B4010570) and by Soonchunhyang Research Fund.

Disclosure of potential conflicts of interest: The authors declare that there is no conflict of interests regarding the publication of this paper.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahn, H. S.; Jung, B. K.; Park, J. R. (2014): Effect of reagents on optical properties of asbestos and remote spectral sensing. *Journal of Convergence*, vol. 5, no. 3, pp. 9-12.

Bolhuis, S.; Kromme, J. (2016): Features of R. http://www.theanalyticslab.nl.

Chung, H; Nah, Y. (2016): Effects of hypervisor on distributed big data processing in virtualizated cluster environment. *KIISE Transactions on Computing Practices*, vol. 22, no. 2, pp. 89-94.

Derdour, Y.; Kechar, B.; Fayçal-Khelfi, M. (2016): Using mobile data collectors to enhance energy efficiency and reliability in delay tolerant wireless sensor networks. *Journal of Information Processing Systems*, vol. 12, no. 2, pp. 275-294.

Finogeev, A. G.; Parygin, D. S.; Finogeev, A. A. (2017): The convergence computing model for big sensor data mining and knowledge discovery. *Human-Centric Computing and Information Sciences*, vol. 7, no. 11, pp. 1-16.

Franceschi, F.; Cobo, M.; Figueredo, M. (2018): Discovering relationships and forecasting PM10 and PM2. 5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. *Atmospheric Pollution Research*, vol. 9, no. 5, pp. 912-922.

Jang, S. J. (2007): *Design and Implementation of Courseware for Learning Normalization and Query Optimization in Database (Ph.D. Thesis).* University of Hanyang, Korea.

Kang, W. M.; Moon, S. Y.; Park, J. H. (2017): An enhanced security framework for home appliances in smart home. *Human-Centric Computing and Information Sciences,* vol. 7, no. 6. pp. 1-12.

Kim, J. W.; Yun, Y. D.; Jung, Y. J.; Kim, K. Y. (2016): Korean collective intelligence in sharing economy using R programming. *Journal of Internet Computing and Services*, vol. 17, no. 5, pp. 151-160.

Li, J.; Wang, L. (2017): The research of PM2.5 concentrations model based on regression calculation model. *American Institute of Physics Conference*, vol. 1794, no. 030005, pp. 1-7.

Li, Y.; Kim, D.; Shin, B. S. (2016): Geohashed spatial index method for a location-aware WBAN data monitoring system based on NoSQL. *Journal of Information Processing Systems*, vol. 12, no. 2, pp. 263-274.

Li, Y.; Li, J.; Chen, J.; Lu, M.; Li, C. (2018): Seed selection for data offloading based on social and interest graphs. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 571-587.

Ratliff, J. K.; Balise, R.; Veeravagu, A.; Cole, T. S.; Cheng, I. et al. (2016): Predicting occurrence of spine surgery complications using "big data" modeling of an administrative claims database. *The Journal of Bone and Joint Surgery*, vol. 98, no. 10, pp. 824-834.

Yoon, M. Y. (2013): Analysis and implications of big data promotion strategy in major countries. *Science and Technology Policy*, vol. 23, no. 3, pp. 31-43.