A Study of the Correlation Between Livestock Data Analysis and the Concentration of PM2.5 - Using the Cloud Computing Platform

Chien-Yuan Tseng[1] and Jui-Hung Chang[2(✉)]

[1] Computer and Network Center, National Cheng Kung University, Tainan 701, Taiwan
P76031488@mail.ncku.edu.tw
[2] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701, Taiwan
changrh@mail.ncku.edu.tw

Abstract. The subject of air pollution is paid increasing attention to in recent years. NH3 has significant effect on PM2.5, especially due to animal excrements and chemical fertilizer. Some PM2.5 monitoring data in Taiwan show that the concentration in south central Taiwan is apparently on the high side, and south central Taiwan has dense livestock and poultry. Therefore, this paper combines the livestock and poultry data opened by Taiwan Council of Agriculture (COA) with the PM2.5 data obtained by air monitoring stations opened by Taiwan Environmental Protection Administration (EPA), and uses SpatialHadoop to build a Cloud platform to analyze the correlation between Taiwan's livestock data and the concentration of PM2.5. The analysis results show that the annual mean concentration of PM2.5 of the air monitoring station in the livestock and poultry dense region is higher than that in other regions by 33%.

Keywords: PM2.5 · Cloud platform · SpatialHadoop

1 Introduction

The cause of fine particulate matter is very complex, it is one of the hazards to the air and environment, it receives close attention of various countries' governments and research units in recent years. The PM2.5 is too small to be blocked by vibrissae in the nasal cavity and general masks, it has significant effect on human respiratory system, cardiovascular and nervous systems. This paper designs a cloud computing platform based on the relationship between Taiwan's livestock data analysis and PM2.5 for data operation. The purpose of Cloud platform is to use Taiwan's livestock data to analyze the correlation with the concentration of PM2.5, the back end uses SpatialHadoop [8] as cloud computing architecture. As the initial data range of this paper is only the data from Taiwan, other countries' data can be added in the future. The display terminal displays the relevance between the location of Taiwan's air monitoring station and the livestock and poultry quantity data of nearby townships on Google Maps.

In 2014, Beijing Municipal Environmental Protection Bureau declared at the Sino-American Engineering Technology Forum that Beijing implemented field

J.-H. Lee and S. Pack (Eds.): QShine 2016, LNICST 199, pp. 75–85, 2017.
DOI: 10.1007/978-3-319-60717-7_8

monitoring of NH3 emission sources, such as chicken farms, cattle farms and pig farms. It was indicated that the NH3 emission was paid little attention to, now it is recognized as a key factor in the formation of PM2.5. It is found that over 80% of NH3 emission came from agricultural fertilizer application [3].

The news of Beijing municipal government monitoring NH3 enlightened the preliminary conception of this paper. Late studies show that the NH3 has significant effect on the concentration of PM2.5, and the NH3 is mostly derived from animal excrements and chemical fertilizer. Therefore, this paper analyzes the correlation between the livestock data and the concentration data of PM2.5, hoping to provide useful findings to control the NH3 emission from livestock farms effectively, and the concentration of PM2.5 can be reduced, so as to reduce the probability of common people's respiratory and cardiovascular diseases.

This paper obtained the livestock and poultry statistics opened by Taiwan COA, including the animal varieties and size of animal of 198 townships, with the locations of 76 air monitoring stations in Taiwan opened by EPA and the concentration data of PM2.5 monitored automatically per hour by the air monitoring stations. This paper uses highly expandable and geographic operation supporting SpatialHadoop to build a Cloud storage computing platform, the map linear distance from the air monitoring station to the nearby township center is calculated, and the daily mean, monthly mean and annual mean concentrations of PM2.5 of 76 air monitoring stations in 2014 are calculated, the results are visualized on Google Maps, the process architecture will be detailed in Chap. 3. Finally, this paper uses Livestock Geographic (LG) algorithm to prove the effect of the quantity of pigs raised nearby the air monitoring station on the concentration of PM2.5 of monitoring station.

Main contributions of this paper:

(1) The Cloud storage computing platform process is created by SpatialHadoop - analyzing livestock data and the concentration data of PM2.5.
(2) The correlation between the livestock data and the concentration data of PM2.5 is analyzed, the result is visualized on Google Maps for research units or experts and scholars to make further analytic investigation.
(3) This paper calculates the six counties and cities with dense livestock and poultry in Taiwan, the annual mean concentration of PM2.5 of air monitoring stations of Yunlin County, Changhua County, Pingtung County, Tainan City, Chiayi County and Kaohsiung City is compared with the annual mean concentration of PM2.5 of air monitoring stations of other counties and cities.
(4) The LG algorithm is used for validation and the experimental results prove that the quantity of pigs raised in counties, cities and townships has substantive effect on the concentration of PM2.5.

The rest part of this paper is arranged as follows, Sect. 2 introduces related technologies and background. Section 3 describes the Cloud platform system architecture. Section 4 introduces the system equipment implementation method, and proposes experimental results. Finally, the contribution of this paper is summarized in Sect. 5.

2 Related Work

In recent years, the application of Geographic Information System (GIS) is increasingly diversified and important, it is used by government agencies for decision making system applications. At present, the whole world pays increasing attention to big data analysis, and the combination of big data analysis result and GIS receives increasing attention of coherent units and academia. Therefore, the cross-domain combined system is a new subject, for example, applications of biomedical image domain [9], and the NASA satellite imagery analysis [11] uses geographic data and big data analysis to obtain useful information. General server architecture does not have very good effectiveness on processing massive data like geographic data. Therefore, how to process and analyze geographic big data efficiently is an important domain in recent years. This chapter discusses and describes related studies, for example, MapReduce-based geographic system, the geographic algorithm and so on, which are popular research areas.

2.1 SpatialHadoop

The cloud computing is more and more important, and in the trend of big data analysis, the MapReduce-based Hadoop big data analysis is well accepted by academia and circles. However, Hadoop has not provided Spatial Index for geographic operation, so the effectiveness of Hadoop on geographic data operation query fails to meet the anticipation (Fig. 1).

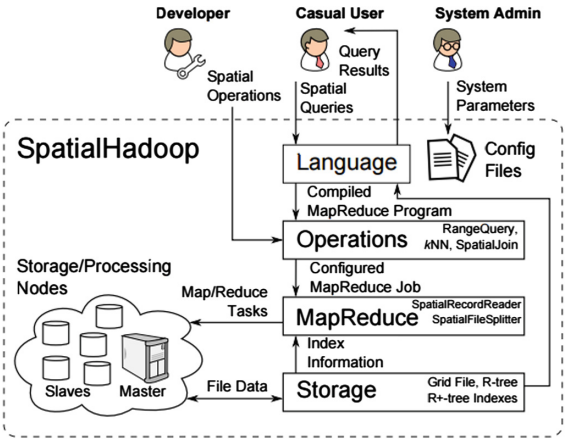

Fig. 1. SpatialHadoop architecture diagram.

SpatialHadoop [10] is an Open Source software developed by Eldawy and Mokbel from the Department of Computer Science & Information Engineering, University of Minnesota. SpatialHadoop uses Spatial Index (Global Index, Local Index) function. The modes of Spatial Index include Grid File [14] and R-tree [12].

2.2 Studies About PM2.5

This paper studied many reports and theses on PM2.5 in relation to NH_3, Chinese reportage describes NH_3 as a neglected cause of PM2.5, it has been noticed for only a few years [4, 7]. Some indicate that NH_3 is likely to change into nitrogen pentoxide and nitric acid (NOx) which are likely to polymerize into aerosol and PM2.5 in the atmosphere. In the final PM2.5 in the atmosphere, 15–35% of nitrogen element is derived from NH_3. In comparison to NOx, the NH_3 is not paid attention to, it may because the previous computation model underestimates the effect of NH_3, the corrected conclusion is that the effect of NH_3 on atmospheric pollution like PM2.5 is non negligible [13]. Some studies indicate that the NH_3 can take part in and accelerate the formation of ammonium bisulfate and ammonium sulfate in the atmosphere directly, it is a key factor in the formation of fine particulate matter. The NH_3 is one of precursors of PM2.5 [15].

3 Cloud Platform System Framework

The Cloud platform system framework is shown in Fig. 2. This paper proposes using SpatialHadoop to build the Cloud platform. The correlation between the livestock data analysis and the concentration of PM2.5 is validated by statistics and the self-developed LG algorithm. The computing result is saved in Hadoop Distributed File System (HDFS) and database. The visualization mode is that the webpage and Google Maps display latitude and longitude coordinates and charts for experts and scholars to research and analyze. The present data are limited to Taiwan, if there are more massive data for operation, SpatailHadoop distributed architecture provides good expansion and elongation for other research units' reference.

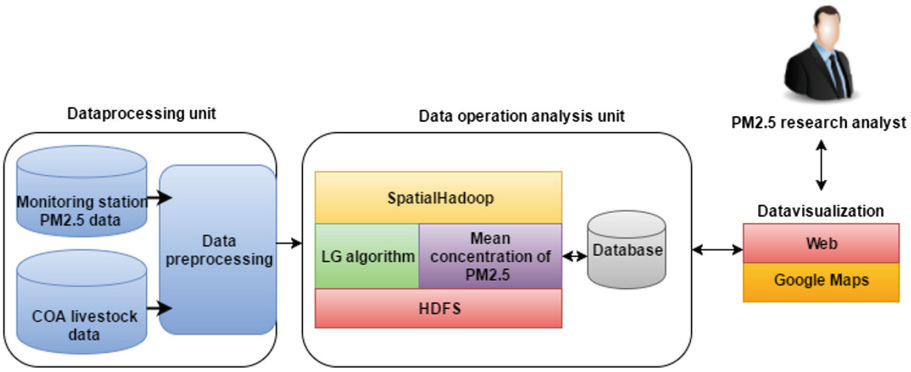

Fig. 2. Cloud platform architecture diagram.

3.1 Data Processing Unit

The data studied by this system are divided into livestock data opened by COA [1] and air monitoring station data opened by EPA [5], preprocessed by different procedures.

1. Livestock data opened by COA:
   This paper extracts data from the hog population investigation report of 2014, the statistical findings of livestock and poultry include the number of livestock farms and livestock and poultry calculated by 198 townships of Taiwan, the statistical livestock and poultry data include pigs, cattle and chickens. This paper uses the latitude and longitude coordinates of township center point to replace the livestock farm address, Taiwan township center point is converted into latitude and longitude coordinate data by the system call geo information graph data cloud service platform (TGOS) [6] and Google Maps. The livestock data established in this paper include id, county, city, township, year, livestock and poultry varieties of the township, the quantity of livestock and poultry of the township, the latitude and longitude coordinates of township center point.
2. Air monitoring station data opened by EPA:
   The gas concentration monitored automatically per hour by 76 air monitoring stations in Taiwan is provided by Taiwan EPA, referring to U.S. Environmental Protection Agency, Taiwan EPA uses the linear regression equation (relationship) of automatic monitoring station and manual monitoring station data to correct the values of automatic monitoring, so as to guarantee the correctness of values. This paper extracts the gas concentration data of 2014, including PM10, NO, NO_2, NO_x, O_3, SO_2, wind direction, humidity and so on. The air monitoring station data include date, monitoring station name, monitoring item, monitoring station address and gas concentration monitored per hour. The uncertain values are planned to be deleted, such as Null, so as to guarantee the correctness of mean value. This paper calls geo information graph data cloud service platform (TGOS) [6] to convert the air monitoring station address into latitude and longitude coordinate data.

3.2 Data Operation Analysis Unit

This paper proposes using SpatialHadoop to create the Cloud platform operation analysis unit. The SpatialHadoop distributed geographic algorithm process is shown in Fig. 3. Uncorrelated data don't enter the Map stage by Spatial Index, the system computing speed is increased. The proposal programs are divided into two types

1. The PM2.5 gas concentration data monitored per hour by 76 air monitoring stations in Taiwan provided by EPA are calculated, the daily mean, monthly mean and annual mean concentrations of PM2.5 are calculated. The way is to use the longitude and latitude of the monitor station to make Grid File grouping. The same PM2.5 gas concentration data from the same monitor station will be distributed to the same group. The total of 76 monitor stations will be divided into 76 groups. When the daily average of PM2.5 is calculated, the key in the stage of Map is set with the designated number of the monitor stations, which the number starts

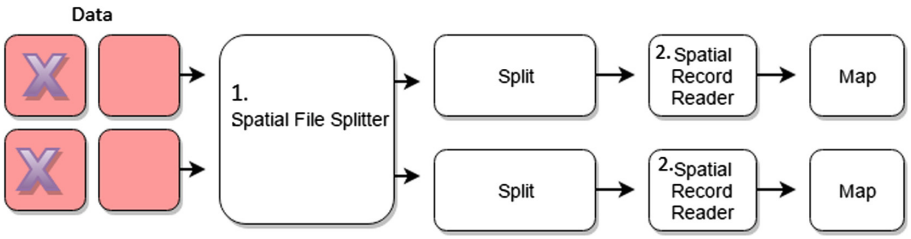

Fig. 3. Cloud platform computing process.

from 1 to 76, and its corresponding value is the monitored concentration data of PM2.5. Therefore, at the stage of reduce, the daily average PM2.5 concentration of every monitor station can be obtained from the calculated result, which is the total of PM2.5 concentration of every monitor station divided by the number of the day's effective data. Similarly, it can also calculate the annual average and monthly average in the same way.

If the average PM2.5 concentration of a specific monitoring station is needed to be calculated, the cloud platform utilizes Spatial File Splitter (Fig. 3.1) and Spatial Record Reader (Fig. 3.2) in the structure of SpatiaHadoop. Before the stage of Map, the PM2.5 concentration data of the specific monitoring station the file in Hadoop Distributed File System (HDFS)needs to be placed into the Map. The unnecessary data from other monitoring stations will be eliminated. Therefore, it can greatly boost up the speed of calculating the specific average PM2.5 concentration data in the cloud platform.

2. Livestock Geographic (LG) algorithm, the substantive effect of pig dense region on the concentration of PM2.5 monitored by nearby air monitoring station is validated, the schematic of the algorithm is shown in Fig. 4. The latitude and longitude coordinates of air monitoring station are used as center of circle, a circle in radius of K km is taken, the quantity of pigs raised in the townships in this circle is calculated.

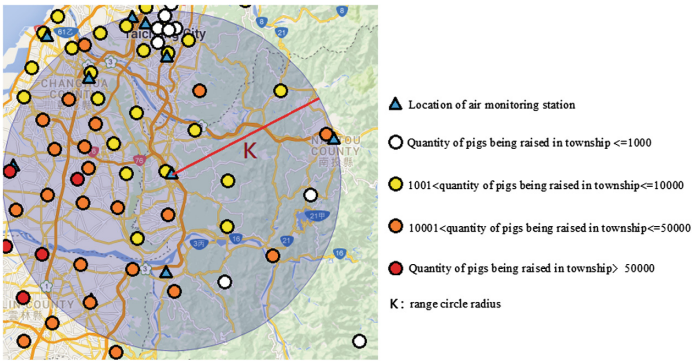

Fig. 4. LG algorithm illustration.

The quantity of pigs raised in the townships within K km around the air monitoring station is calculated by LG algorithm. The correlation between the quantity of pigs raised nearby and the change in the concentration of PM2.5 of air monitoring station is investigated, there are detailed experimental results in the experiment chapter, the algorithm is described below:

```
Input:
      k: The range circle radius is set as 30 km in this
      paper.
      n: Number of air monitoring stations, set as 76 in
      this paper.
      m: Number of townships, set as 198 in this paper.
      Latitude and longitude coordinates of n air
      monitoring stations set S1={{longitude 1, latitude
      1)…{longitude n, latitude n)}.
      Latitude and longitude coordinates of m township
      center points set S2={(longitude 1, latitude
      1)…(longitude m, latitude m)}.
      Quantity of pigs raised in m townships set S3
      ={(quantity of pigs raised 1)…(quantity of pigs
      raised m)}.
Output:
      Sum: quantity of pigs raised in townships in the
      circle in radius of K km of 76 air monitoring
      stations.
program LG
  for i = 1 to n do
      sum[i]=0
      for j = 1 to m do
            If (distance(locationi, locationj) < k) then
                  sum[i] += S3[j]
            end if
      end for
  end for
  return sum
```

The distance equation of Geographic Median algorithm is Eq. (1) Spherical law of cosines [2]:

$$\text{distance} = \arcsin(\sin(\text{lat1}) * \sin(\text{lat2}) + \cos(\text{lat1}) * \cos(\text{lat2}) * \cos(\text{lon2} - \text{lon1})) \quad (1)$$

4 Experimental Results

The computer equipments of SpatialHadoop distributed architecture for this paper are two sets of Intel I7-4770 CPU, 8 GB RAM, 1 TB hard disk, the database server equipment is Intel I5-4590 CPU, 12 GB RAM, 1 TB hard disk, the software is Microsoft SQL Server 2012.

4.1 Analysis of Concentration of PM2.5 in Counties and Cities with Dense Livestock and Poultry in Taiwan

According to the hog population investigation report in the end of July 2014 provided by COA and the livestock and poultry statistics of the third quarter of 2014, the livestock and poultry data of 19 counties and cities and 3 off islands of Taiwan are calculated. There were 8,198 pig farms and 5,539,130 pigs, 2,244 cattle farms and 145,877 cattle, 5,760 chicken farms and 92,142,813 chickens in Taiwan (including off islands) in 2014. Figure 5 shows the top six counties and cities with the highest quantitative proportion of pigs, cattle and chickens in Taiwan are Yunlin County, Changhua County, Pingtung County, Tainan City, Chiayi County and Kaohsiung City.

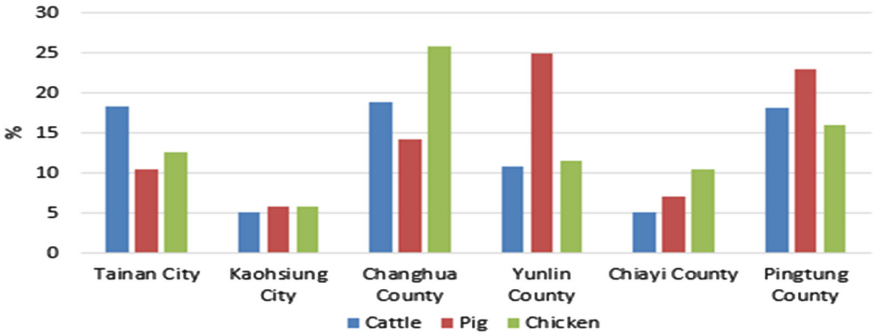

Fig. 5. Proportions of pigs, cattle and chickens in livestock data of various counties and cities of Taiwan.

This paper uses SpatialHadoop architecture to calculate the annual mean concentration of PM2.5 in 2014 of 76 air monitoring stations of Taiwan, and compares the annual mean concentration of PM2.5 in Yunlin County, Changhua County, Pingtung County, Tainan City, Chiayi County and Kaohsiung City with other counties and cities, the annual mean concentration of PM2.5 of monitoring stations in the counties and cities with dense livestock farms is 30.57 μg/m^3, the annual mean concentration of PM2.5 of monitoring stations in other counties and cities is 23.17 μg/m^3, as shown in Fig. 6.

4.2 Analysis of Quantity of Pigs Raised in the Range of Air Monitoring Station

The analysis in Sect. 4.1 combines the annual mean concentration of PM2.5 of air monitoring stations with the livestock data of the counties and cities where the monitoring stations are, the quantity of pigs raised in the counties and cities nearby the air monitoring stations is not considered. For example, Changhua County is to the west of Nantou County, Yunlin County is to the southwest, the proportion of pigs in Nantou County is not high, but the quantity of pigs raised in nearby counties and cities is very

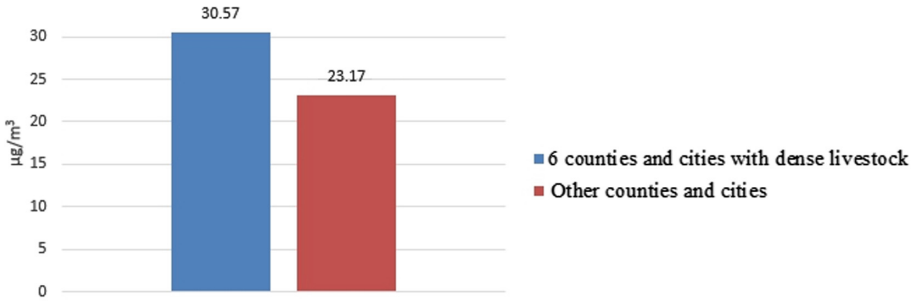

Fig. 6. Annual mean concentration of PM2.5 of monitoring stations in six counties and cities with high proportion of pig, cattle and chicken and other monitoring stations in livestock data.

large, this is not seen in the statistical data in Sect. 4.1. Therefore, this paper uses LG algorithm to calculate the quantity of pigs raised in the townships within the range circle in radius of 30 km of 76 air monitoring stations in whole Taiwan, the quantities of pigs raised nearby the air monitoring stations are ranked, the annual mean PM2.5 of Nos. 1 to 38 air monitoring stations and Nos. 39 to 76 air monitoring stations is calculated, as shown in Fig. 7, the average annual mean concentration of PM2.5 in the first 38 townships with pigs nearby the air monitoring stations is 30.46 μg/m^3, the average annual mean concentration of PM2.5 of the last 38 monitoring stations is 21.49 μg/m^3.

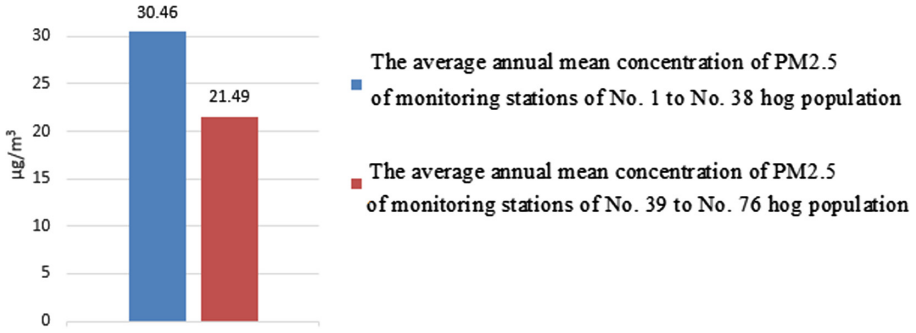

Fig. 7. Annual mean concentration of PM2.5 of the first 38 and the last 38 hog population in townships nearby monitoring stations.

5 Conclusion

This paper uses SpatialHadoop to build a Cloud platform for analyzing the correlation between livestock data and annual mean concentration of PM2.5 of air monitoring stations. The results are displayed on Google Maps. This paper designs two experiments to validate the effect of quantity of livestock and poultry on the annual mean concentration of PM2.5, the results are:

1. The average annual mean PM2.5 of monitoring stations in six counties and cities with dense livestock and poultry is compared with that of monitoring stations in other counties and cities, which are 30.57 μg/m^3 and 23.17 μg/m^3 respectively.
2. The LG algorithm is used to calculate the annual mean PM2.5 of No. 1 to No. 38 air monitoring stations and No. 39 to No. 76 air monitoring stations, which is 30.46 μg/m^3 and 21.49 μg/m^3 respectively.

There are two verification modes, the results show that the annual mean concentration of PM2.5 monitored by the air monitoring station in the livestock and poultry dense region is higher than the annual mean PM2.5 in the nondense regions by about 33%.

The concentration of PM2.5 may be influenced by many other factors, which are not considered in the research scope of this paper, such as the factory effluence, automobile and motorcycle emissions and natural environment factors. The future research will analyze related data to increase the accuracy. In addition, this paper creates an analysis module first, so only the data of 2014 are learned about, the historical data will be analyzed in the future, the ultimate objective is to use the Cloud platform to work out why Taiwan's annual mean concentration of PM2.5 is relatively high for related studies' reference.

References

1. http://agrstat.coa.gov.tw/sdweb/public/book/Book.aspx
2. https://en.wikipedia.org/wiki/Spherical_law_of_cosines
3. http://hk.on.cc/cn/bkn/cnt/news/20140529/bkncn-20140529051643333-0529_05011_001_cn.html
4. http://news.qq.com/a/20150301/022140.htm
5. http://taqm.epa.gov.tw/taqm/en/YearlyDataDownload.aspx
6. http://tgos.nat.gov.tw/tgos/web/tgos_home.aspx
7. http://scitech.people.com.cn/BIG5/n1/2016/0222/c1007-28138290.html
8. http://spatialhadoop.cs.umn.edu
9. Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop GIS: a high performance spatial data warehousing system over MapReduce. Proc. VLDB Endow. **6**(11), 1009–1020 (2013)
10. Eldawy, A., Mokbel, M.F.: A demonstration of SpatialHadoop: an efficient MapReduce framework for spatial data. Proc. VLDB Endow. **6**(12), 1230–1233 (2013)
11. Eldawy, A., Alharthi, S., Alzaidy, A., Daghistani, A., Ghani, S., Basalamah, S., Mokbel, M. F.: A demonstration of SHAHED: a MapReduce-based system for querying and visualizing satellite data. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 2015
12. Guttman, A.: R-trees: a dynamic index structure for spatial searching. SIGMOD Rec. **14**(2), 47–57 (1984)
13. Heald, C.L., Collett Jr., J.L., Lee, T., Benedict, K.B., Schwandner, F.M., Li, Y., Clarisse, L., Hurtmans, D.R., Van Damme, M., Clerbaux, C., Coheur, P.-F., Philip, S., Martin, R.V., Pye, H.O.T.: Atmospheric ammonia and particulate inorganic nitrogen over the United States. Atmos. Chem. Phys. **12**, 10295–10312 (2012)

14. Nievergelt, J., Hinterberger, H., Sevcik, K.C.: The grid file: an adaptable, symmetric multikey file structure. ACM Trans. Database Syst. **9**(1), 38–71 (1984)
15. Li, L., Kumar, M., Zhu, C., Zhong, J., Francisco, J.S., Zeng, X.C.: Near-barrierless ammonium bisulfate formation via a loop- structure promoted proton-transfer mechanism on the surface of water. J. Am. Chem. Soc. **138**, 1816–1819 (2016)