Int. J. Advance Soft Compu. Appl, Vol. 9, No. 2, July 2017
ISSN 2074-8523

Air Quality Analysis Based On MapReduce and K-Means: A Decision Making System

Abderrahmane SADIQ [1], Abdelaziz EL FAZZIKI [1], Jamal OUARZAZI [2], Mohamed SADGAL [1]

[1] Computer Systems Engineering Laboratory, Cadi Ayyad University of Marrakech, Morocco
abderrahmane.sadiq@edu.uca.ma, {elfazziki, sadgal}@uca.ma
[2] Laboratoire Physico-Chimie des Matériaux et Environnement (LPCME) URAC 20, Cadi Ayyad University of Marrakech, Morocco
ouarzazi@uca.ma

Abstract

Urban air pollution management requires an advanced modeling and information analyzing and processing techniques. Designing a system of air quality management must be based on a distributed and adaptive problem-solving approach, which is the aim of this work. In this article, we infer real-time and detailed information regarding the air quality throughout the city of Marrakech based on pollutants and meteorological data (historical and real-time) reported by the existing monitoring stations. All these types of inputs fed the system, which uses an advanced computational module for analyzing and extracting needed information through the use of a set of algorithms and online analytical processing (OLAP) tools. The advanced features of an air quality management system are necessary and expected to be helpful to stakeholders, planners and decision makers, so that they can reliably generate a statement and prevision on the air quality, simulate and analyze more information in the decision making process. The proposed system in this work will provide an intelligible tool for the air quality data collection and analysis, based on the use of K-means algorithm and Hadoop framework: HBase for data storage and MapReduce process for data processing.

Keywords: *Air quality analysis, Hadoop, K-means, MapReduce, OLAP.*

1 Introduction

With the continuous development of society, the effectiveness of human production has been drastically improved and the emergence of new scientific products make human's life better than before. However, the continuous increases in productivity bring also damage to the environment, due to the various factory emissions and vehicle exhausts and other pollution sources. It has been found out that this phenomenon is particularly serious in developing countries. Public health, the environment and living conditions, depend closely on it; the real attention that must be carry in the coming decades is decisive. Protecting the air is an environmental issue that affects policies as diverse as energy, transport and territories use planning. This problem can be controlled by measuring, monitoring, alert forecasting, awareness and action in the context of a supported scientific research.

An air quality management system is a tool for improving the air quality. It must be based on the new standards in the quality of ambient air and help in establishing the foundation for air quality management of the outside air [1]. It is a framework for managing air areas within cities and regions, which can allow appropriate measures to specific atmospheric emissions in a particular city or region. Also, it defines regional airshed, which allows coordination when air pollution crosses regions and contains requirements related to industrial emissions [2]. This system should allow a better collaboration of stakeholders to reduce emissions in the transportation sector [3]. The most common functions of such systems are: Editing reports, online analytical processing, data mining (DM) and predictive analytics. Furthermore, the solution to growing volumes of data that demands fast and effective retrieval of information lie in integrating the principles of data mining over a distributed environment managed by Big Data tools like Apache Hadoop [4]. For this, we propose the use of the K-mean algorithm which is the well-known and commonly used clustering method algorithm [20] over a Hadoop MapReduce process [5][6] to make the clustering method applicable to large scale data [7] and gives a great flexibility and speed to execute the algorithm over a distributed framework [8].

In this work, we aim to design an HBase [4] based data warehouse and discusses the use of the proposed approach for the air quality management decision support system (DSS) development. This approach is illustrated over a few sections starting with a brief literature review of the existing studies on air quality and pollution management systems and the use of Big Data tools in DSS. In section 3 a case study is presented. Section 4 is devoted to the Hadoop HBase based air quality data storage design. Section 5 will be dedicated to the K-Means Algorithm based on Hadoop MapReduce. This will be followed by the experimental results and a conclusion in section 7.

2 Related Work

Decision support systems enhancement using Hadoop as a data hub to optimize the decision making infrastructure is a new emerging strategy. Many research works have proposed a method to leverage the Hadoop framework by effectively integrating it to the existing data warehouse such as [9] that proposed a study on big data integration with data warehouse built using relational technology mainly for operational sources. In the same context, Facebook, for example, developed Hive, an open source data warehousing solution built on top of Hadoop [10]. It is based on familiar concepts of tables, columns and partitions, providing a high-level query tool for accessing data from existing warehouses [9].

Concerning the management and assessment of air quality using decision support system, many works have been proposed, such as [6] that describe the structural components of such an integrated system, designed to give a support in decision making, connected with forecasting and urban air quality managing. Such system ensures the assessment of air pollution and allows predicting the air quality in diverse urban situations. According to Larssen [10] in order to provide scientifically noise and decision relevant information supporting planning and management of the air quality, such air quality management system is needed in order to select the right decisions for the protection of human health and the ecosystem from an increasing impact of air pollution.

Many projects have also addressed the issue of air quality data integration; like Appetise project [11] that aims to produce a database containing pollution data combined with other related data such as weather records, and to develop tools for analyzing and visualizing this data. The Time Map project [12] has also developed data analysis software that allows visualization of distributed spatiotemporal data sets, and interactive maps. These solutions can be improved by using big data tools as proposed in this work, in order to enhance the system performance and ensure a great flexibility and speed and make needed algorithms applicable to large scale data. In recent studies such as [13], the authors have addressed the use of a big data framework for air quality data analysis and proposed various solutions. They propose an indoor air quality analysis based on Hadoop and the use of a K-means clustering algorithm with MapReduce in order to analyze the time series of pollutants concentration and the relationship between the carbon dioxide (CO_2) concentration and the number of people in the study area.

3 The Case Study

The objective of this case study is to develop an air quality management system for air areas within the city of Marrakech, which can allow appropriate measures to specific atmospheric emissions in particular areas [14].

Marrakech-Tensift-El Haouz region is situated in central Morocco. It has a population of 3,102,652 (According to the 2004 census) and covers an area of 31,160 km². The city of Marrakech is not an industrial character marked city, but it suffers the effects of pollutants produced by vehicle exhaust systems. The main sources of air pollution are then anthropogenic ones and particularly coming from auto motor vehicles.

This study is based on stations that provide information's and measures of air pollutants concentration and has been carried out in order to implement a larger project aiming cadastre development and modeling of the air quality in the agglomeration. The city of Marrakech has three stations, two of which Daoudiat and Jamaa El Fna (JEF) are located in the center of Marrakech whereas Mhamid is located south west in an area where traffic is less dense. Fig. 1 shows the location of each station (Jamaa EL Fna [A], Mhamid [B], and Daoudiat [C]); other details are presented in Table 1.

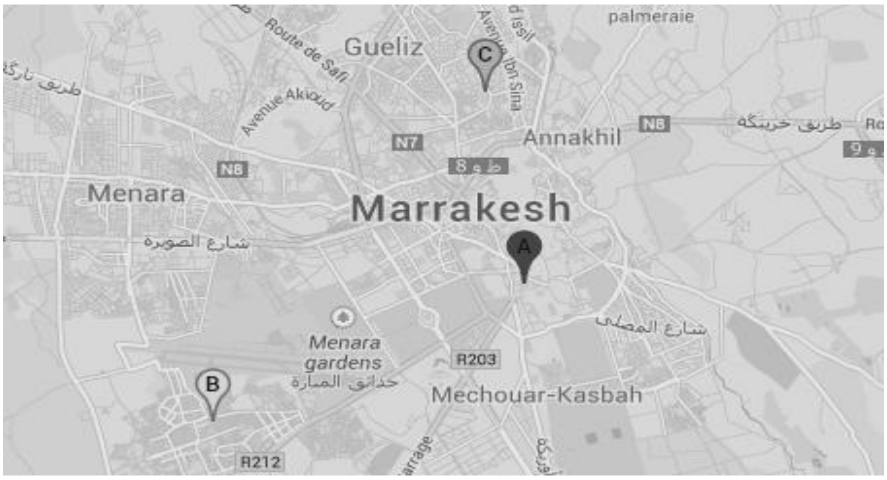


Fig. 1: Marrakech's air quality monitoring stations (By Google Maps)

Table 1: Features of the permanent air quality monitoring stations in Marrakech

Station	Mhamid	JEF	Daoudiat
Type	Urban	Urban	Urban
Latitude (N)	31.596289	31.620254	31.653676
Longitude (W)	8.043691	7.988897	7.995688
Start date	06/01/2009	06/01/2009	03/01/2010

The study focuses on the following pollutants: Sulfur Dioxide (SO_2), Nitrogen dioxides (NO_2), Carbon Monoxide (CO), Particulate Matter (PM10), and Ozone (O_3). In Morocco, the National Meteorology Direction (NMD) retrieves and manages data from station measurements concerning these pollutants. The data to integrate in the system is available for the three stations dating from 2009-2011.

4 HBase Based Data Storage

4.1 Data storage design

Each air quality monitoring station periodically sends gathered data to the system, causing a large amount of records every day issued from the three stations sensors. However, the traditional data processing method compresses data sequence by sampling, and analyzes and forecasts the trend with partial data.

Therefore, it presents many difficulties related to data gathering and processing and storage growth, since the system database will receive millions of data records after running a long time. This method is also designed based on several prerequisites and desired questions in advance and sampling results are only used to reply to some questions [9].

New methods and technologies for the collection and storage of data have appeared with the development of sensors and big data [15]. They analyze the relevant information from different views, which will solve many other problems while avoiding doing random sampling that can only address some questions. Thus Hadoop HBase is chosen for the increasing data management and storage [16]. HBase is a Java open source; non-relational, distributed database modeled after Google's BigTable [4] [15]. It is a database with high reliability, high performance, column storage, scalable characteristics based on the Hadoop distributed file system (HDFS). Its goal is the hosting of very large tables with billions of rows and millions of columns atop clusters of commodity hardware [13]. An HBase table is organized as key-value and each table contain a series of row records [16]. The Rowkey is the unique identifier of a row in the table. A {row, column, version} tuple exactly specifies a cell in HBase and a cell content is interpreted bytes.

As most of column-oriented databases, HBase is structured into a set of tables composed of a set of rows and whose physical storage is organized by groups of columns called column families; a column family can contain a very large number of columns. For each row, a column exists if it contains a corresponding value. In this study the pollutants concentrations and other data are continuously collected by air monitoring station's sensor system. All the data are extracted and stored in the HBase. The logical star schema of the air quality data to store in the HBase tables is given below (Fig. 2) and table 1illustrates an HBase table structuring.

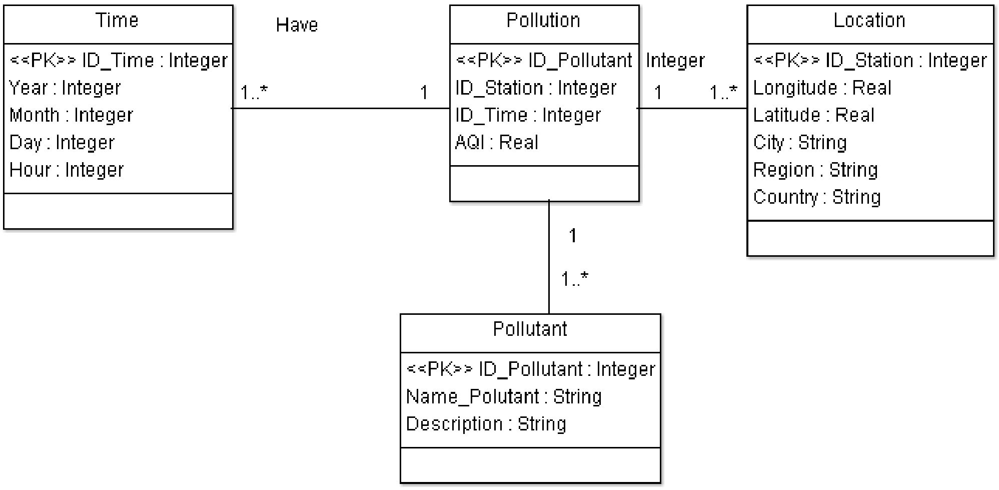


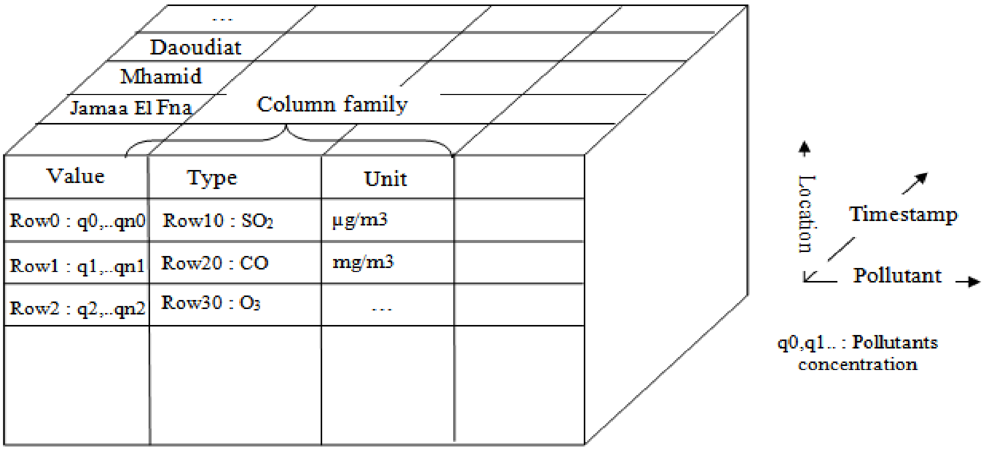
Fig. 2: Multidimensional conceptual schema of our example, pollution data is presented according to pollutant, records time and location

Table 2: The ozone (O_3) table in HBase

Row Key	Pollution (Column family)		Location(Column family)				Time				Pollutant		
	API	…	Zone	Region	City	…	Hour	Day	Month	..	Name	Hourly -Max	…

4.2 Multidimensional data presentation

For each pollutant, a daily sub-index of air quality is calculated and stored, based on gathered data. The stored data are processed to produce data on a particular location for a particular pollutant. This data are then loaded into a multidimensional OLAP cube [17] that gives possibility to run analysis queries on a big amount of data depending on selected dimensions and represent data along some measure of interest. This vision corresponds to a structuring of data analysis along several axes (or dimensions) [3]. The data cube structuring for the air quality management is shown in Fig. 3.



Value	Type	Unit	
Row0 : q0,..qn0	Row10 : SO_2	µg/m3	
Row1 : q1,..qn1	Row20 : CO	mg/m3	
Row2 : q2,..qn2	Row30 : O_3	...	

Fig. 3: Presentation of the air quality monitoring data in a multidimensional cube

5 Hadoop MapReduce and K-means

5.1 Hadoop MapReduce

The parallel computing programming model MapReduce is used to process data collected by sensors. MapReduce is proposed by Google Labs [8]. It is a programming model for processing large data sets and easily writing applications which process vast amounts of data in-parallel on large clusters. It is used to distribute computing, on clusters of computers. The MapReduce job divides the input dataset into independent fragments which are processed by the map tasks in a parallel manner. The framework sorts the outputs of the maps and then input to the reduce job. In our case, both the input and the output of the job are stored in a Hadoop HBase [16] [18].

5.2 K-means clustering using MapReduce

K-means is one of the simplest algorithms that solve clustering problems. The procedure follows a simple way to classify a dataset through a several number of clusters [7]. This algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster centroid or center of gravity.

The K-means algorithm is based on few steps:

- Randomly selects k objects from the whole objects which represent initial cluster centers.
- The remaining objects are assigned to the most similar cluster, based on the distance between each object and the cluster center (centroids).
- A new mean for each cluster is then calculated.
- The process is then iterated until the criterion function converges.

The definition of the input and the output of the implementation is the first step in designing the MapReduce routines for K-means. The input data vectors and the initial set of chosen cluster centers is stored in the input directory of the HDFS. It is given as a <key,value> pair, where ‘key’ is the cluster center and ‘value’ is the serializable implementation of vector in the data set. The Map and Reduce routines need also an output directory to house the result of clustered data [18].

Once the initial set of clusters with their centers and the data to be clustered are properly organized in the input directory and loaded on the master node then the clustering algorithm can be accomplished.

The Map routine is then executed in order to form the ‘key’ field in the <key,value> pair. The Mapper routine contains the required instructions to compute the distance between a given data set and clusters center. It is structured in such a way that it computes the distance between the vector value and each of the cluster centers mentioned in the cluster set and simultaneously keeping track of the cluster to which the given vector is closest. Once the computation of distances is complete the vector should be assigned to the nearest cluster [19].

Once the Mapper is invoked the given vector is assigned to the cluster that it is closest related to. After the assignment is done the centroid of that particular cluster is recalculated. The recalculation is done by the Reduce routine and also it restructures the cluster to prevent creations of clusters with extreme sizes, i.e. cluster having too less data vectors or a cluster having too many data vectors.

Based on the iterations and the value of ‘K’ the resulting clusters of data and the updated centroids can be found in the results directory which can be echoed to the output directory in HDFS [18]. The Figure 4 below illustrates the MapReduce based iterative K-Means.

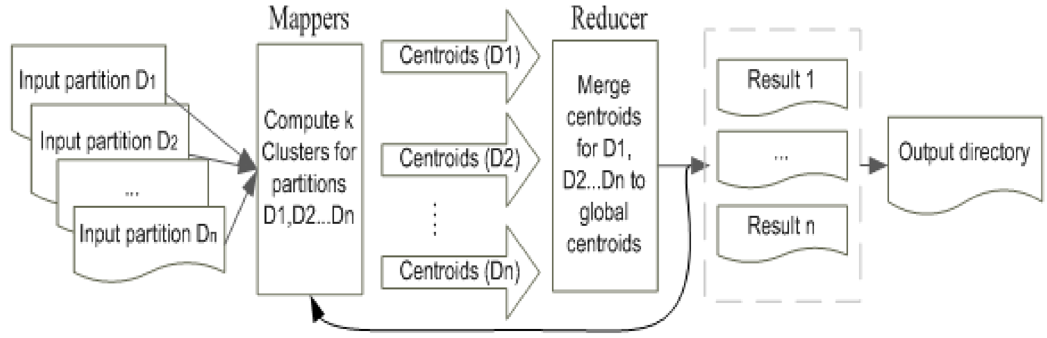


Fig. 4: The architecture of the online based K-means Map-Reduce

We design the MapReduce Based K-means classes by following the algorithm discussed below [19].

Algorithm 1: Mapper design for K-Means Clustering

```
procedure KmeanMapper
LOAD Cluster file
mp = MapFile
Create two list
listNew = IistOld
CALL read (MapClusterFile)
newmp = MapClustcr()
dv =0
Assign correct centeroid
read(dv)
calculate centerorid
dv = minCenter
CALL KmeansReduce()
end procedur=0
```

Algorithm 2: Reducer design for K-Means Clustering

```
procedure KmeanReducer
NEW LisOfCluslers
COMBINE resultant clusters from MAP CLASS.
if cluster size too high or too low then RESIZE the cluster
CMax = findMaxSize(ListofClusters)
CMin = findMinSize (ListofCIusters)
if CMax>1/20 totalSize then ResizeCluster)
WRITE cIuster FILE to output DIRECTORY
```

Algorithm 3: Implementing K-Means Function

```
procedure KMEANS FUNCTION
if Initial Iteration then LOAD cluster file from DlRECTORY
else READ cluster file from previous iteration
Create new JOB
SET MAPPER to map class defined
SET REDUCER to reduce class define paths for output directory
SUBMIT JOB
```

6 Experimental Results

6.1 K-means based MapReduce

Ozone is a secondary pollutant; it is the result of a photochemical reaction between the volatile organic compounds and nitrogen oxides in the presence of solar radiation. The dataset which contains few years' data are analyzed in the experiment.

The study was performed on hourly measurements using two variables: the ozone concentration and solar radiation. The data are derived from the Jamaa El Fna (JEF) permanent measurement station of two years 2009 and 2010. Partition of results into three clusters seems to be the most obvious.

The first cluster belongs to a part of the day from 18h to 21h where sunlight is variable depending on the season: A not null solar radiation during summer, nonexistent during the winter and half sunny in the spring. The second cluster corresponds to the time period from 8h to 18h, which is a sunny day period throughout the year (non-zero solar radiation) where ozone production takes place. The third cluster is spread over the period from 21h to 8h after which solar radiation is null and then there is no ozone production.

Tables 3 and 4 below shows the k-means parameters and the resulting cluster's data.

Table 3: K-means used parameters

K-Means parameters	
Clusters	3
Max Iteration	10
Distance normalization	variance
Average computation	McQueen

Table 4: Clusters centroids

Attribute	Cluster 1	Cluster 2	Cluster 3
O_3 JEF	59,836678	68,937295	25,737021

The following Figure 5 shows the resulting clusters.

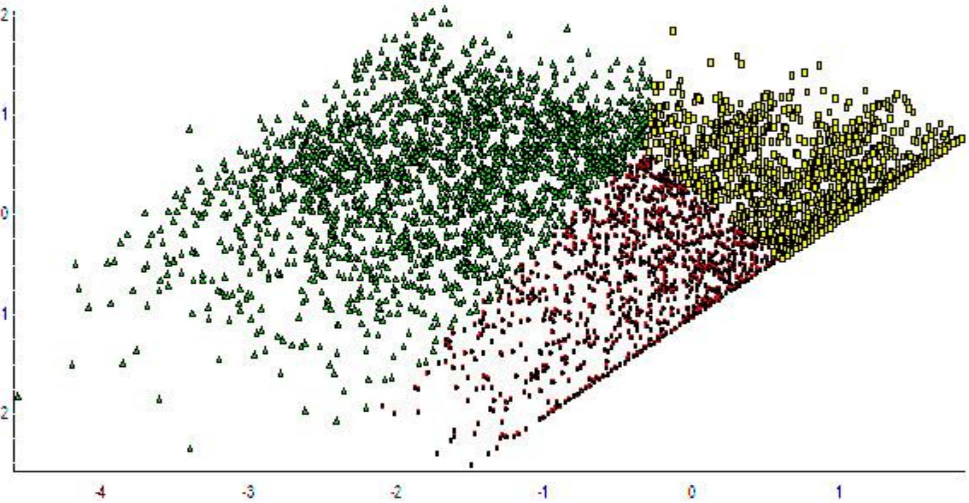


Fig. 5: The analysis resulting three clusters

The Figures 6 and 7 below shows the Ozone concentration difference during winter and summer, which explain the difference between the resulting clusters, caused by the ozone production rate which depends on solar radiation.

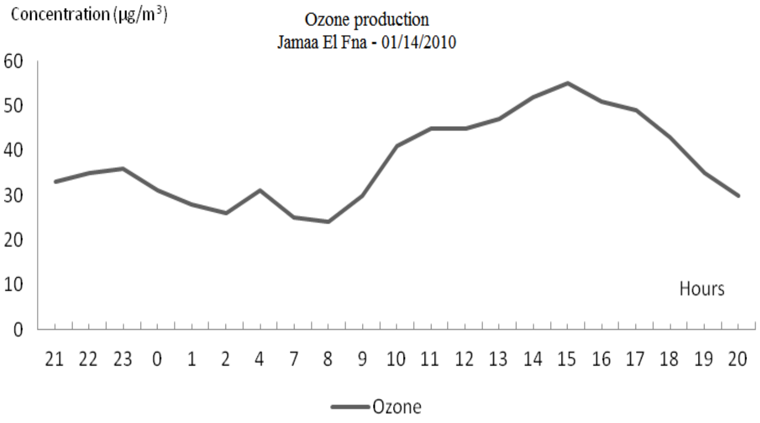


Fig. 6: Ozone production profile within 14/01/2010 in Jamaa El Fna station (one day data)

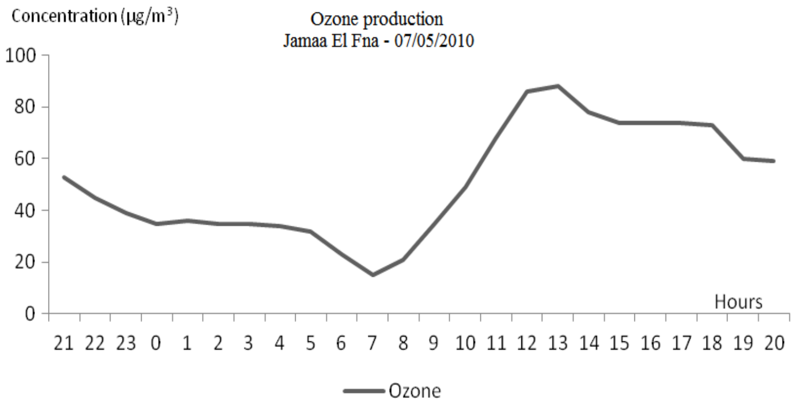


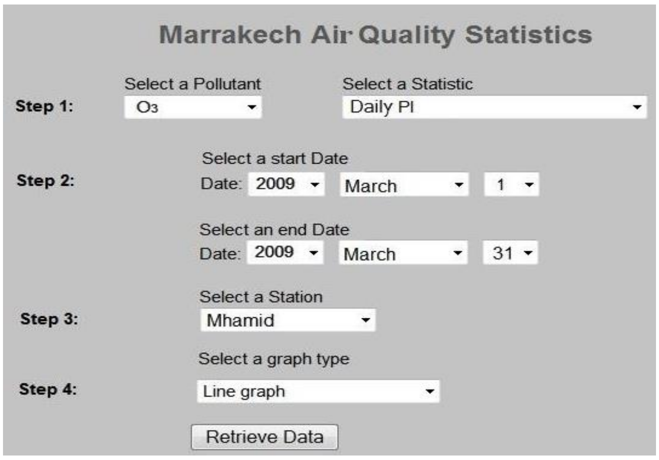
Fig. 7: Ozone production profile within 05/07/2010 in Jamaa El Fna station (one day data)

6.2 Air quality monitoring

It is important to maintain the end user interface of an analysis tool, user friendly where data analysis tasks are simpler to perform. The system generates the analysis queries based on selected options from an easy to use user interface.

The purpose is to convert the air quality data into valuable information by applying quick and effective analysis, creating different representations of this data and providing pertinent and in-time responses to specific users' queries. This can provide an overview for decision makers to study the evolution of air pollutant levels and a support when defining actions to manage the air quality in the study area.

In this study, the required air quality information can be extracted by defining the categories in a search form, e.g. station name, year, pollutant, start date, end date, output type, etc. A few screen shots from the air quality management system for Marrakech's user interface are presented below (Fig. 8-10).


Marrakech Air Quality Statistics

Step 1: Select a Pollutant: O3 — Select a Statistic: Daily PI

Step 2: Select a start Date — Date: 2009 March 1

Select an end Date — Date: 2009 March 31

Step 3: Select a Station: Mhamid

Step 4: Select a graph type: Line graph

Figure 8: The user interface selected options to display the Ozone PI in Mhamid station

Fig. 8 shows the user interface to select the desired output; in this case we have chosen to display the calculated Ozone (O_3) pollution index values during March 2009 in Mhamid station.

The corresponding OLAP queries are generated automatically based on the selected options in the user interface. Fig. 9 and 10 show examples of the resulting graphs. Fig. 10 presents the variation in ozone level during two weeks (Data gathered from Mhamid station), compared to solar radiations.

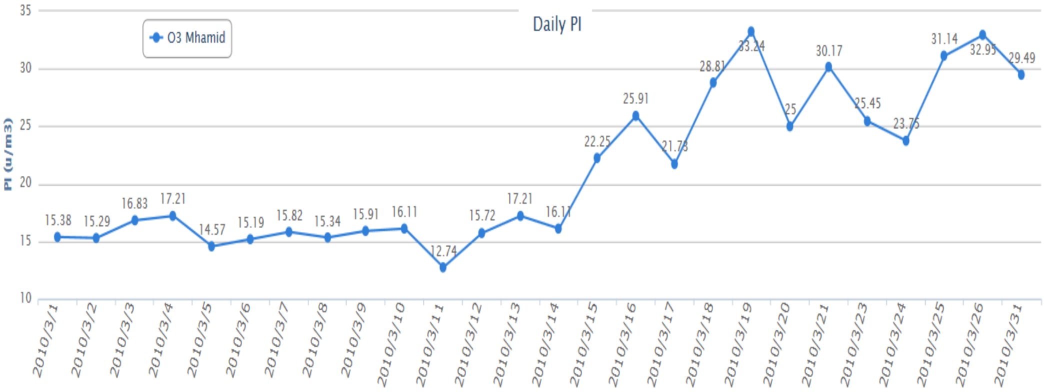


Figure 9: The Ozone PI in Mhamid station results

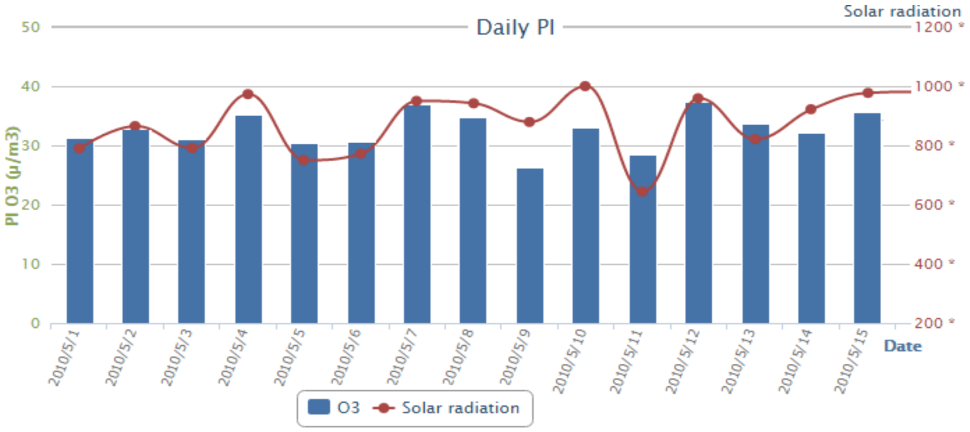


Figure 10: Evolution of the Ozone level in Mhamid station according to solar radiations

7 Conclusion

The growth in the volume of information exchange in the world engages huge quantity of data processing. Most of the organization, including millions of customers needs a system to process a big amount of data daily. Such system must meet the following requirement [9]:

- Fast data loading;
- Fast query processing;
- Highly efficient storage utilization;

In order to satisfy this big data challenges in terms of fast processing and managing highly varying and big amount of data, a data warehouse must be integrated with the Hadoop engine to exploit its MapReduce parallel processing power.

We have, through this paper discussed the implementation of the K-means clustering algorithm over distributed data gathered from different air quality monitoring stations and managed using a Hadoop based system. This solution provides a robust and efficient system for processing big volumes of data.

The case study is successfully performed. The data granularity is up to the hour and the generated reports include certain effects, such as the station position within the city and sunlight effect. In the study, all time series data are stored in an HBase and the K-means clustering algorithm based on MapReduce is used to analyze this data. According to the classification results of ozone concentration, It can be found easily that the ozone production average depends on the seasons and the solar radiation during each day whence the K-means classification into three clusters.

Acknowledgements

The authors wish to acknowledge the contributions of other members of computer systems engineering laboratory and physical-chemistry of materials and environment laboratory (URAC 20) for their helpful discussions and the availability of all resources that have helped make this work in the best conditions. They would also like to thank the reviewers for their remarks and suggestions.

References

[1] Sivertsen, B. 2012. Air quality management planning, *Chemical Industry & Chemical Engineering*, Vol.18, No.4, 667-674.

[2] Li, S., Chou, S., and Pan, S. 2000. Multi-resolution spatiotemporal data mining for the study of air pollutant regionalization, In *Proceedings of the 33rd Hawaiian International Conference on System Sciences*, Island of Maui, Hawaii.

[3] Muhammad, S. 2010. Development and implementation of air quality data mart for ontario, canada: A case study of air quality in Ontario using OLAP tool. Master Thesis. Centre for Geographical Information Systems: Lund University. Available at : https://lup.lub.lu.se/student-papers/search/publication/3559141.

[4] Apache Hadoop Documentation. 2014. available at: http://hadoop.apache.org, last visited 11 June 2015.

[5] Dean J. and Ghemawat S. 2008. MapReduce: simplified data processing on large clusters, *Communications of the ACM,* Vol.51, No.1, 107-113.

[6] Ekanayake, J., Li, H., Zhang, B., Gunarathne, T., Bae, S., Qiu, J., and Fox, G. 2010. Twister: A runtime for iterative mapreduce, In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, Chicago, US, 810-818.

[7] George, A. 2013. Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm, *International Arab Journal of Information Technology*, Vol.10, No.5, 467-476.

[8] Lammel, R. 2008. Google’s MapReduce Programming Model – Revisited, *Science of Computer Programming*, Vol. 70, 1-30.

[9] Das, K. and Mohapatro, A. 2014. A Study on Big Data Integration with Data Warehouse, *International Journal of Computer Trends and Technology*, Vol.9, No.4, 188-192.

[10] Larssen, S. 2006. Proposed plan for the development of an AQM strategy for Bangladesh, Air quality management (AQM) project, Norwegian Institute for air research.

[11] Matejicek, L. 2005. Spatial modelling of air pollution in urban areas with GIS: a case study on integrated database development, *Advances in geosciences*, Vol.4, 63-68.

[12] Johnson, I. and Wilson, A. 2003. The TimeMap project: developing time-based GIS display for cultural data, *Journal of GIS in Archaeology*, Vol. 1, 124-135.

[13] Wang, T., Sun, Y., Tian, S., Yu, L., and Cui, W. 2014. Indoor air quality analysis based on Hadoop, In *Proceedings of the 35th International Symposium on Remote Sensing of Environment*, Beijing, China.

[14] Richards, M., Ghanem, M., Osmond, M., Guo, Y., and Hassard, J. 2006. Grid-based Analysis of Air Pollution Data, *Ecological Modelling*, Vol. 194, No.1-3, 274–286.

[15] Chang, F., Dean, J., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. 2008. Bigtable: A Distributed Storage System for Structured Data, *ACM Transactions on Computer Systems*, Vol.26, No.2, 1-26.

[16] George, L. 2011. HBase: The Definitive Guide, O'Reilly Media, California.

[17] Wang, B., Gui, H., Roantree, M., and O'Connor, M. 2014. Data cube computational model with Hadoop MapReduce, In *Proceedings of the 10th International Conference on Web Information Systems and Technologies*, Barcelona, Spain.

[18] Zhao, W., Ma, H., and He, Q. 2009. Parallel K-Means Clustering Based on MapReduce, In *Proceedings of the 1st International Conference on Cloud Computing*, Beijing, China, 674-679.

[19] Anchalia, P., Koundinya, A., and Srinath, N. 2013. MapReduce Design of K-Means Clustering Algorithm, In *Proceedings of the International Conference on Information Science and Applications*, Suwon, South Korea, 1-5.