Water Contamination Analysis

Abstract

Water quality is defined in terms of the chemical, physical, and biological content of water. The water quality of rivers and lakes changes with the seasons and geographic areas, even when there is no pollution present. There is no single measure that constitutes good water quality. For instance, water suitable for drinking can be used for irrigation, but water used for irrigation may not meet drinking water guidelines. Water quality guidelines provide basic scientific information about water quality parameters and ecologically relevant toxicological threshold values to protect specific water uses. This paper introduces an efficient method of comparing many different quality parameters affecting water. So, that it become easy to observe the trend in quality parameter that is most influential and needed more attention and also we would be able to monitor whether we were successful in controlling it. The methodology consists of three steps: data collection, aggregate calculation and comparison. The process begins with the collection of data which can be easily procured Online. The entire data is fed into Hadoop and aggregates are easily calculated using the MapReduce facility. After the aggregates are calculated, we graphically compare the different parameters over the four of tenure, using python.

Keywords: MapReduce, Quality parameters, Matplotlib

I. INTRODUCTION

In the last few decades there has been an enormous development in the extent and efficiency of data collection techniques. This sector, concerned with data collection, has been put to use in numerous fields including medicine, military, national development, state characteristics etc. This allows the various agencies to analyze the collected data and make predictions, draw comparisons and make sense out of the raw data.

The importance of analyzing and extracting useful information from large quantity of data is evident from the sheer applicability of data collection.
This project is concerned with the data collected about the list of drinking water quality affected habitations all over India due to and also the districts present within. Information about various attributes is collected. The attributes include State Name,

District Name, Block Name, Panchayat Name, Village Name, Habitation Name, Quality Parameter, Year.

These type of graphs can provide a visual aid to better understand the relative position of the states in various fields. We can easily find which parameter is excessively growing and need more attention.

The data is obtained from the Internet and is processed using MapReduce to simplify the data processing complexity. Finally the simplified data is utilized for the purpose of comparison.

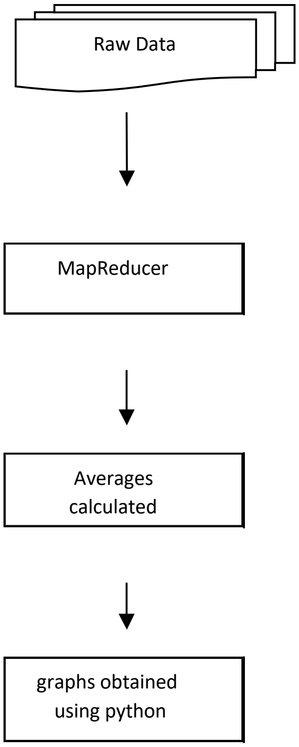
II. MOTIVATION

Water is most essential for livelihoods and for other consumptions. Over-draw and extensive use of pesticides and insecticides for irrigation have made the sources unpotable in many area; excess nitrate in 19387 habitations in 10 states (Rajasthan-7693, Karnataka-4077; Maharashtra-4552).In coastal areas saline water intrusion resulted in contamination of the potable ground water aquifers; 12425 habitations in 15 States (Rajasthan-4428). Therefore, we choose the topic to analyze which element (Fluoride, Arsenic, Iron, etc.) is most responsible for affecting water quality, over the period of 4 year and then graphically represented the output. So, that it is easy to observe the trend in quality parameter that is most influential and needed more attention and also we would be able to monitor whether we were successful in controlling it.

III. METHODOLOGY

The method is composed of three main steps: data collection, applying MapReducer and graph construction.

The basic principle of the proposed method is summarized as follows:



Data is collected from https://data.gov.in/catalog/water-quality-affected-habitations .

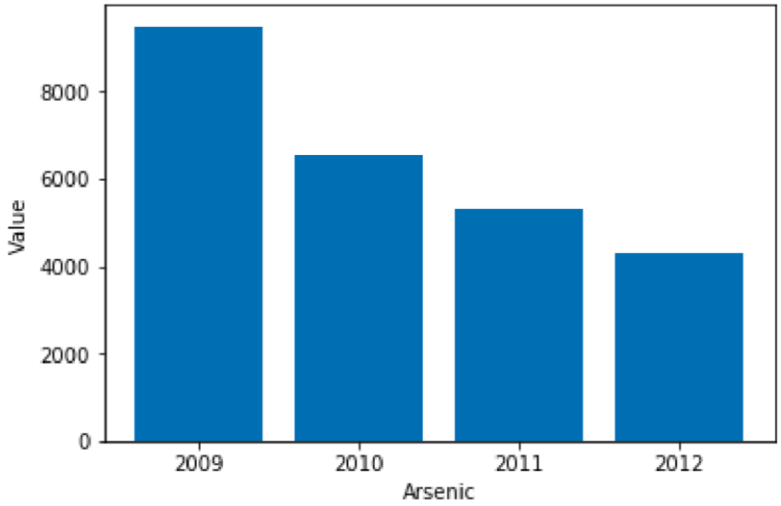
Attributes presents in the data set are State Name, District Name, Block Name, Panchayat Name, Village Name, Habitation Name, Quality Parameter, Year.

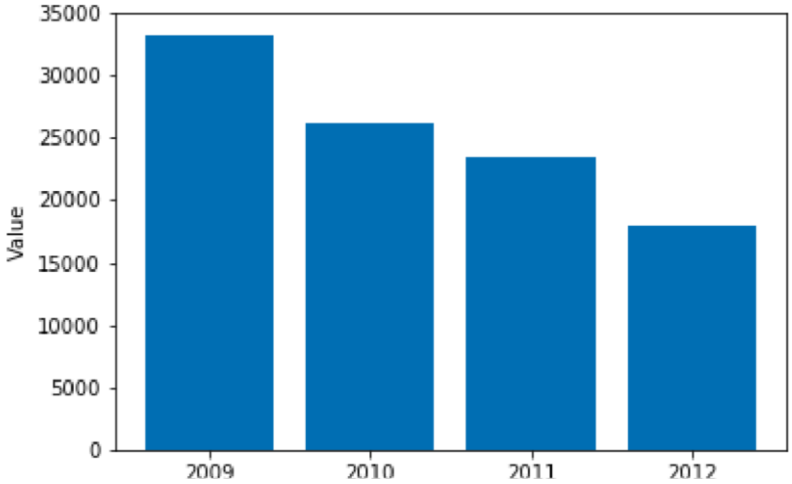
The algorithm used here for aggregation is MapReduce. Generally, MapReduce paradigm is based on sending the computer to where the data resides! MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage. In Map stage, the map or mapper's job is to process the input data. Generally, the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data. Reduce stage, is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS. During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

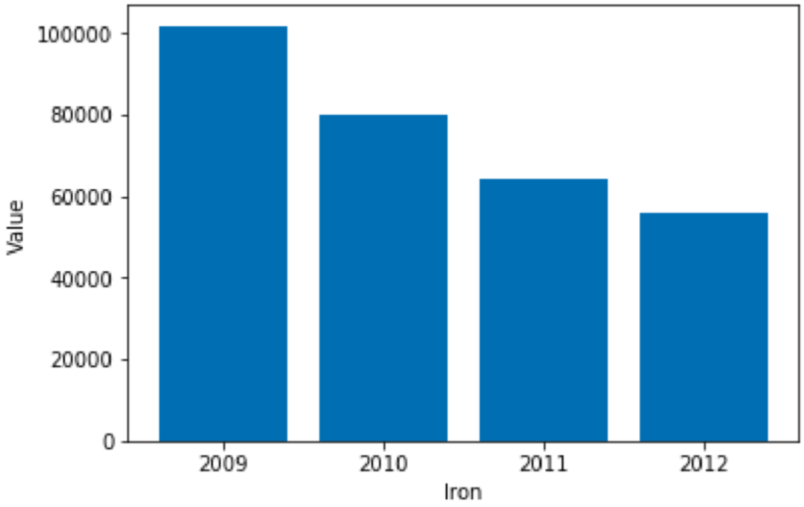
In Mapper, input file is passed to the mapper function line by line.And in Reducer (key, value) attribute for MapReduce will be (Year, Quality Parameter).

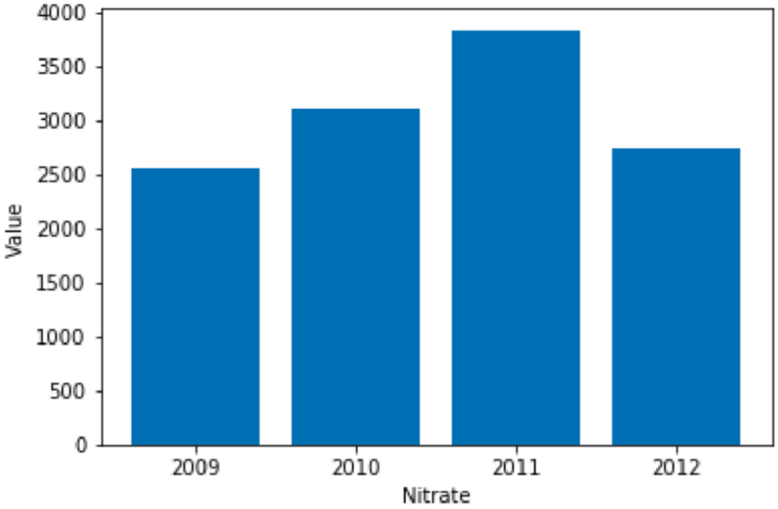
Output from the MapReduce is value of quality parameters responsible for water degradation over four years. Then data Obtained from Output will be represented graphically using Matplotlib (an opensource python library).

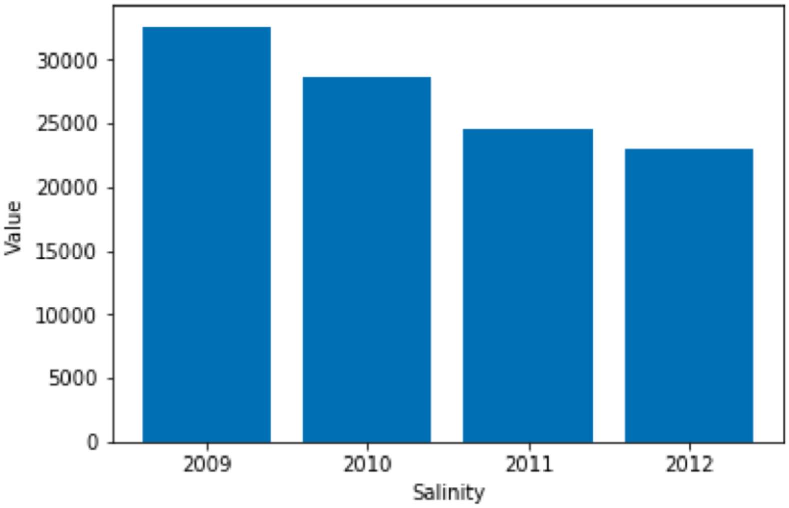
IV. EXPERIMENTAL RESULTS

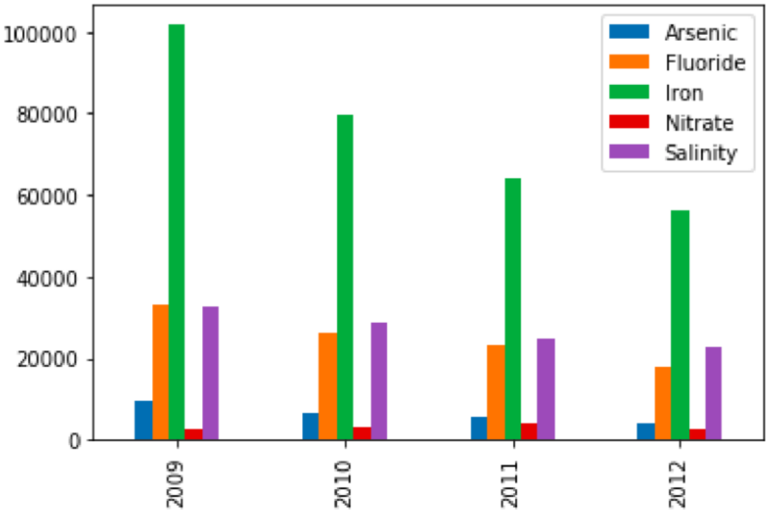


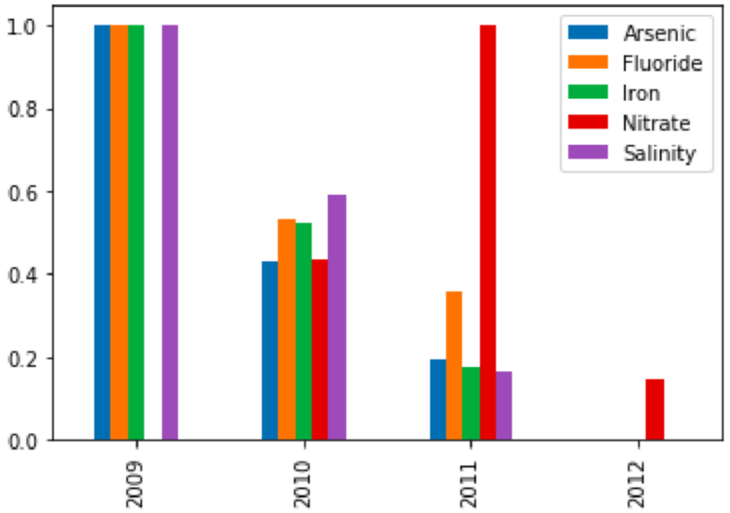





4000
3500
3000
2500
2000
1500
1000
500
0
Value
2009
2010
2011
2012
Nitrate


30000
25000
20000
15000
10000
5000
0
Value
2009
2010
2011
2012
Salinity


100000
80000
60000
40000
20000
0
Arsenic
Fluoride
Iron
Nitrate
Salinity
2009
2010
2011
2012



From the above trend we can observer that in 2009 nitrate was not in the picture but since 2010 it's been extensively used and in 2011 it was ranked 1st parameter to pollute the water. But hopefully, we were able to control it significantly in 2012. And all the other parameters were controlled each successive year successfully.

V. CONCLUSION

This tool can help the government make more informed decisions and formulate plans and policies accordingly. This tool can also be used by other developers to extend this project and utilise the data collected to create other helpful applications. This project can be extrapolated to include real time modification in the data and a mobile application or website can also be developed. We would like to consider these add on as future work.

VI. REFERENCES

[1] https://data.gov.in/catalog/water-quality-affected-habitations .

[2] Google's MapReduce Programming Model—Revisited_Ralf L¨ammel Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[3] Levy E. and Silberschatz A., "Distributed FileSystems: Concepts and Examples" Source from AIAA 2011: Survey of Parallel Data Processing in Context with MapReduce by Madhavi Vaidya

[4] Zhang, Z. Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron. International Journal of Patten Recognition and Artificial Intelligence 13 (6):893-911 (1999).

[5] Review of Distributed File Systems: Concepts and Case Studies ECE 677 Distributed Computing Systems.