Smart Pollution Affected Areas Analysing Technique-A Hadoop Based Approach

N Bharath Kumar
B.E., Under Graduate, Department of Computer Science and Engineering, Misrimal Navajee Munoth Jain Engineering College(Affiliated to Anna University), Chennai, Tamil Nadu, India.
E Janaki
B.E., Under Graduate, Department of Computer Science and Engineering, Sri Muthukumuran Institute of Technology (Affiliated to Anna University),Chennai, Tamil Nadu, India.

Abstract – The critical impact of pollutions on human health and environment in one hand and the complexity of pollutant concentration behavior in the other hand lead the scientists to look for advance techniques for monitoring and predicting the urban air quality, water quality and noise quality. Additionally, recent developments in data measurement techniques have led to collection of various types of data about air quality. Such data is extremely voluminous and to be useful it must be processed at high velocity. Due to the complexity of big data analysis especially for dynamic applications, online forecasting of pollutant concentration trends within a reasonable processing time is still an open problem. The purpose of this paper is to present an online forecasting approach based on Hadoop based approach to analyse the air, water and noise quality. In order to overcome the computational requirements for large-scale data analysis, distributed computing based on the Hadoop platform has been employed to leverage the processing power of multiple processing units. The MapReduce programming model is adopted for massive parallel processing in this study. Based on the online algorithm and Hadoop framework, an online analytical system is designed to analyse the air pollution, water pollution and noise pollution. The results have been assessed on the basis of Processing Time and Efficiency.

1. INTRODUCTION

1.1 BIG DATA

Big data is a term for data sets that are so large or complex that traditional data processing application softwares are inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem."Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research.

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model.

MapReduce is a processing technique and a program model for distributed computing based on java. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

1.2 PROJECT DESCRIPTION

The critical impact of pollution on human health and environment led to look for advance techniques for monitoring and analysing the air, water and noise quality data. The purpose of this project is to analyse and interpret the data of various pollutions and their effects on human health and environment. In order to overcome the computational requirements for large-scale data analysis, distributed computing based on the Hadoop platform has been employed to leverage the processing power of multiple processing units.

2. LITERATURE SURVEY

[1] M.Mazhar Rathore,Awais ahmad,Anand Paul,Seungmin Rho,IoT-based system for smart city development and urban planning using Big Data analytics,vol-2 ,Issue-5,2013.

The rapid growth in the pollution in urban cities demands that services and an infrastructure be provided to meet the needs of city inhabitants.This is IOT-based system for smart city development and urban planning using Big Data analytics. The IoT System consists of various types of sensor deployment, including smart home sensors, vehicular networking, weather and water sensors, air quality sensors, and surveillance objects.The proposed system is implemented using Hadoop with Spark, voltDB, Storm or S4 for real time processing of the IoT data to generate results to establish the smart city.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Client requests for the quality data in a particular area it provides the data which is difficult for a common Most recently designed sensors used in the earth and monitoring stations are generating continuous stream of data. This stream of data is used by various applications to represent those data diagrammatically When the man to understand. There are several embedded devices which record the air, water and noise quality data and provide them to applications for representing them visually.

3.2 PROPOSED SYSTEM

Similar to the existing practice, this system collects the data from the monitoring stations but stores all the data in a particular location (Hadoop Distributed File System). When a client requests for the data of a particular area this system provides the output data in such a way that whether it is harmful or not and also the effects when it is harmful. This system also provides interpreted results for a particular parameter of a pollution. The results can be displayed in a format required by the client i.e either graphically or tabular format.

4. SYSTEM REQUIREMENTS

4.1 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete engineers as the starting point for the system design.

➢ System	: Intel Core i3.
➢ Hard Disk	: 40GB.
➢ Monitor	: 15 VGA Color.
➢ Mouse	: Logitech.
➢ Ram	: 4GB.

SOFTWARE REQUIREMENTS

➢ Operating system	: Ubuntu.
➢ Coding Language	: Java, JSP Servlet.
➢ Front End Tool	: Eclipse Helios.
➢ Database	: HDFS, MySql.
➢ Back End Tool	: Hive ,Sqoop.

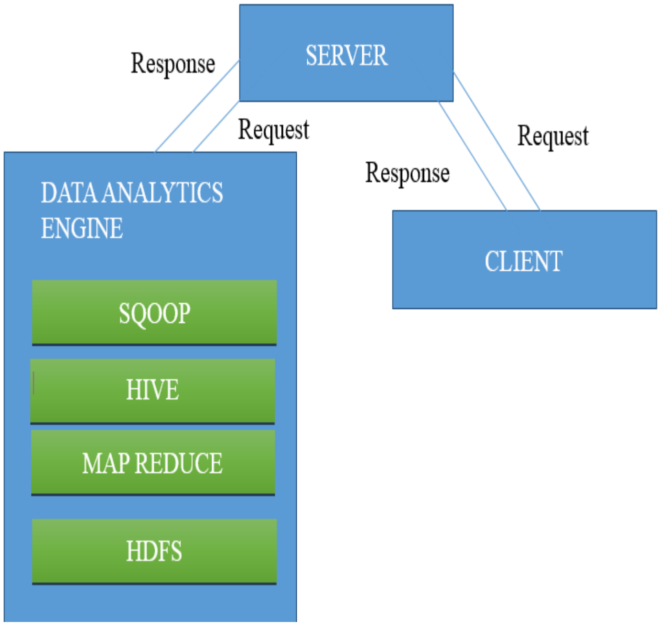
4.3 SUMMARY

In this chapter, the requirements of both hardware and software are discussed to execute the project.

In the next chapter we will discuss about SYSTEM DESIGN.

5. SYSTEM DESIGN

5.1 ARCHITECTURE DIAGRAM



LIST OF MODULES

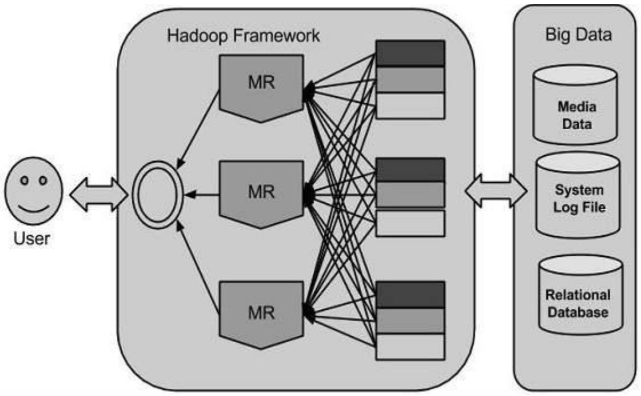
- Server
- Client
- Data Analytics Engine

6. SYSTEM IMPLEMENTATION

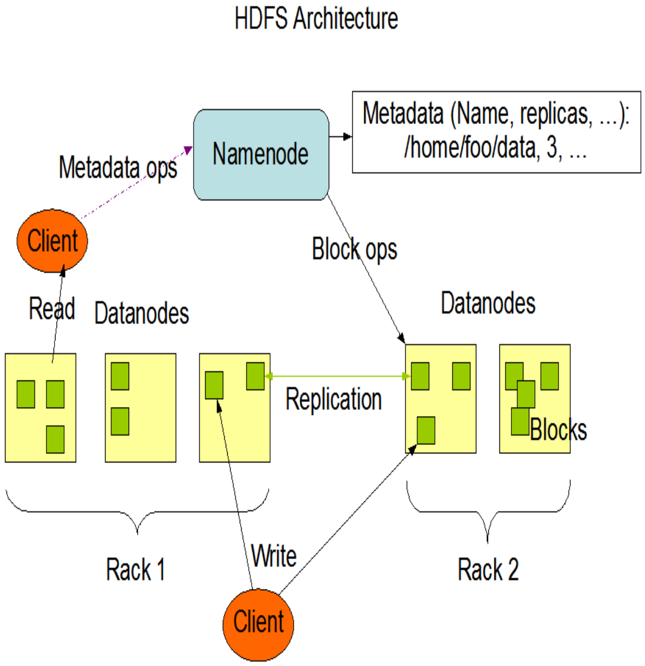
SOFTWARE ENVIRONMENT

6.1 HADOOP

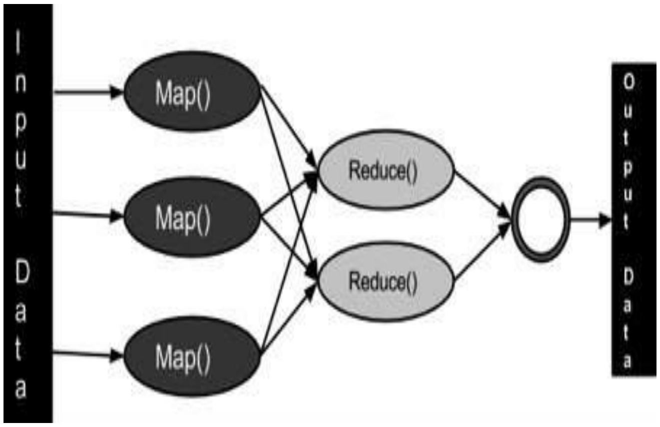
Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.Hadoop makes it possible to run applications on systems with thousands of commodity hardware nodes, and to handle thousands of terabytes of data.



6.2 HDFS



6.3 MAP REDUCE



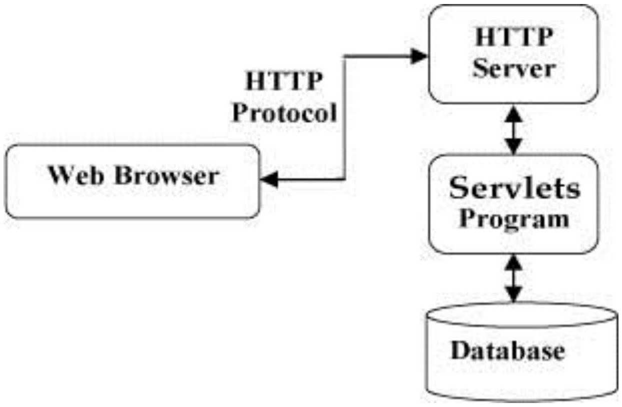
6.4 APACHE HIVE

Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

6.5 APACHE SQOOP

Sqoop is a tool designed to transfer data between Hadoop and relational databases or mainframes. You can use Sqoop to import data from a relational database management system (RDBMS)

6.6 SERVLET ARCHITECTURE



7. SYSTEM TESTING

7.1 SYSTEM TESTING

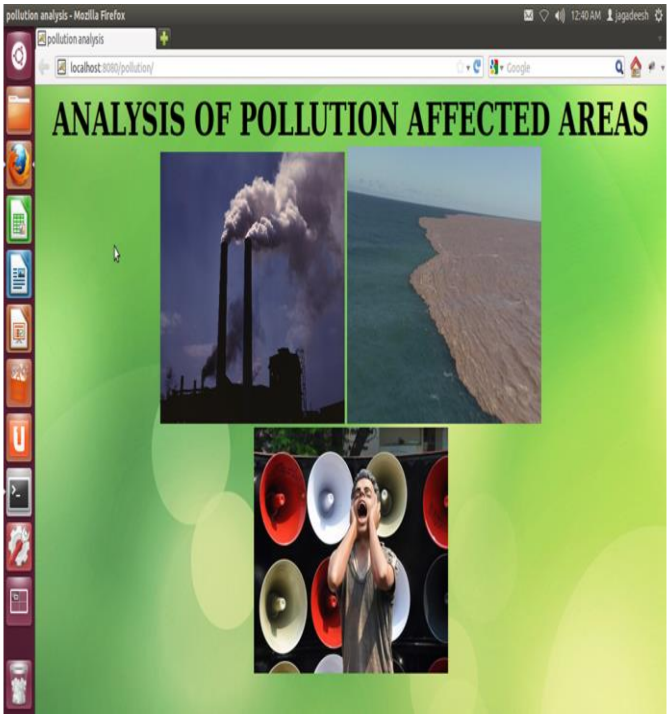
Testing is performed to identify errors. It is used for quality assurance. Testing is an integral part of the entire development and maintenance process. Testing is a set of activities that can be planned in advance and conducted systematically. For this reason a template for software testing, a set of steps into which we can place specific test case design techniques and testing methods should be defines for software process.

8. CONCLUSION

To overcome the most challenging problems of big pollution data, memory usage and computation time, a parallel Hadoop platform was developed based on Hadoop-MapReduce in this study. The proposed system had been utilized for analyse and interpret the data of various pollutions and their effects on human health and environment. The obtained results seemed extremely encouraging and suggested that the proposed system could allow training models for very large scale data set in a reasonable time. Splitting the whole dataset over data nodes and training each subset of data in parallel is the research work. It is also planned to compare the obtained results by employing Hadoop and MapReduce programming with the online algorithm that had been executed on a single machine to evaluate the feasibility of the online distributed system.

9. SAMPLE

A.2.1 MAIN PAGE

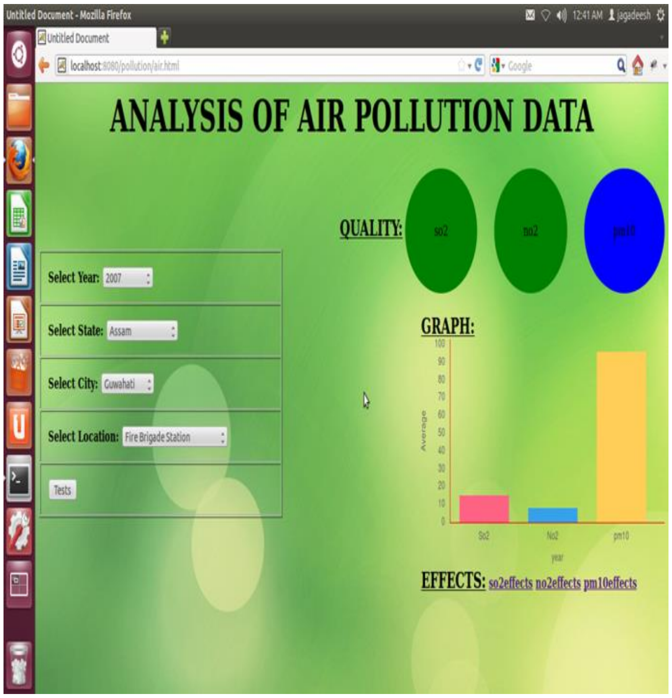


pollution analysis - Mozilla Firefox

localhost:8080/pollution/

ANALYSIS OF POLLUTION AFFECTED AREAS

A.2.2 AIR DATA ANALYSIS



Untitled Document - Mozilla Firefox

localhost:8080/pollution/air.html

ANALYSIS OF AIR POLLUTION DATA

QUALITY: so2 no2 pm10

Select Year: 2007

Select State: Assam

Select City: Guwahati

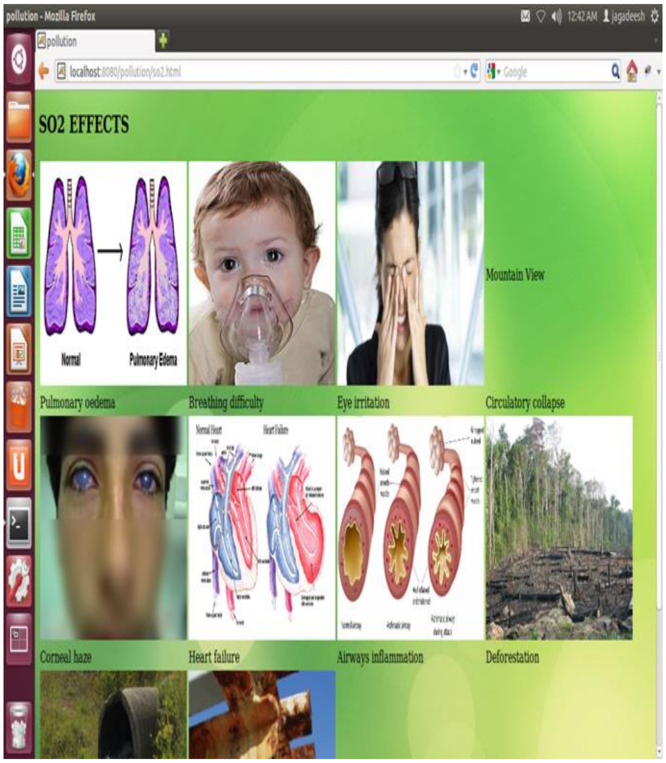
Select Location: Fire Brigade Station

Tests

GRAPH:

EFFECTS: so2effects no2effects pm10effects

A.2.2.1 SO2 EFFECTS



pollution - Mozilla Firefox

localhost:8080/pollution/so2.html

SO2 EFFECTS

Pulmonary oedema

Breathing difficulty

Eye irritation

Circulatory collapse

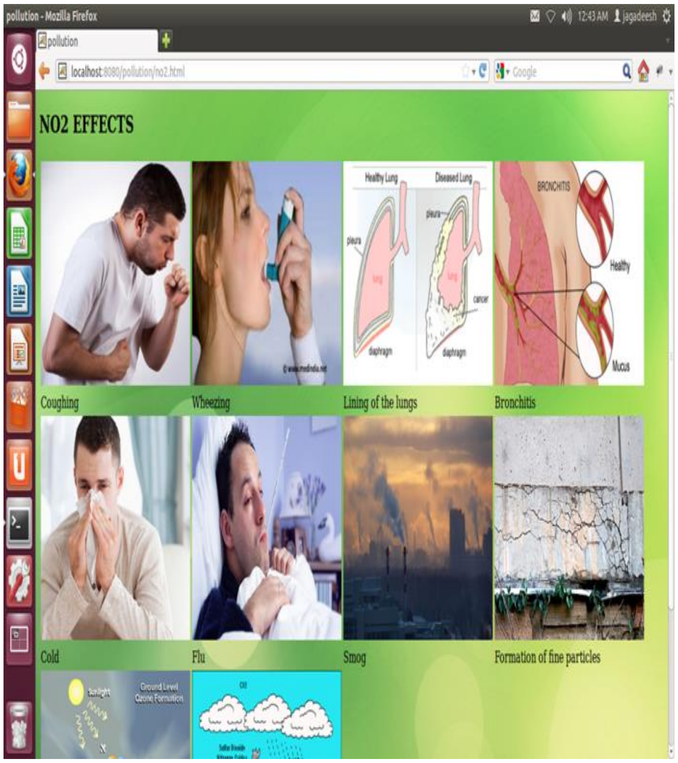
Corneal haze

Heart failure

Airways inflammation

Deforestation

A.2.2.2 NO2 EFFECTS



pollution - Mozilla Firefox

localhost:8080/pollution/no2.html

NO2 EFFECTS

Coughing

Wheezing

Lining of the lungs

Bronchitis

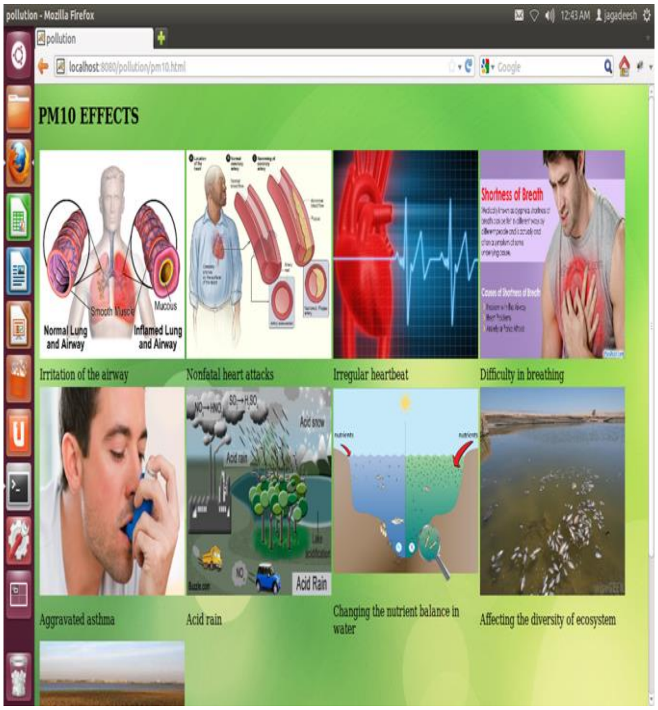
Cold

Flu

Smog

Formation of fine particles

A.2.2.3 PM10 EFFECTS


pollution - Mozilla Firefox

localhost:8080/pollution/pm10.html

PM10 EFFECTS

Irritation of the airway

Nonfatal heart attacks

Irregular heartbeat

Difficulty in breathing

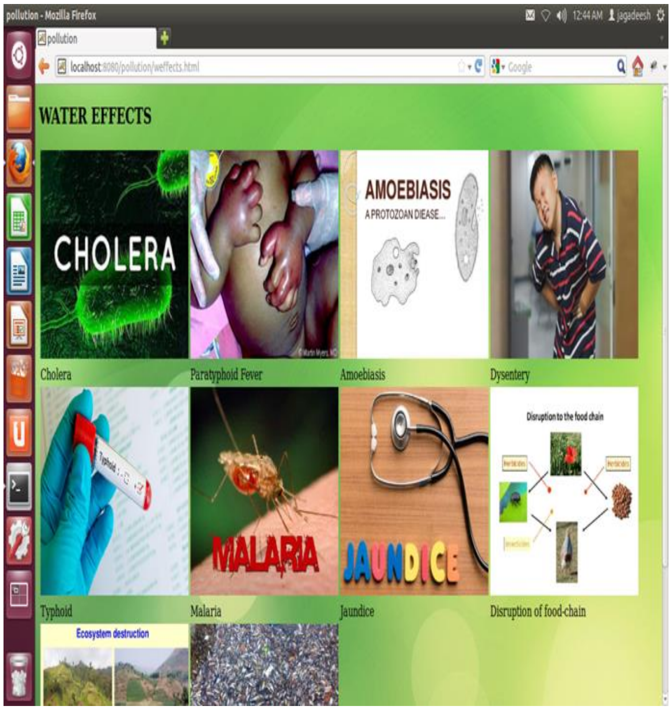
Aggravated asthma

Acid rain

Changing the nutrient balance in water

Affecting the diversity of ecosystem

A.2.3.1 WATER EFFECTS


pollution - Mozilla Firefox

localhost:8080/pollution/weffects.html

WATER EFFECTS

Cholera

Paratyphoid Fever

Amoebiasis

Dysentery

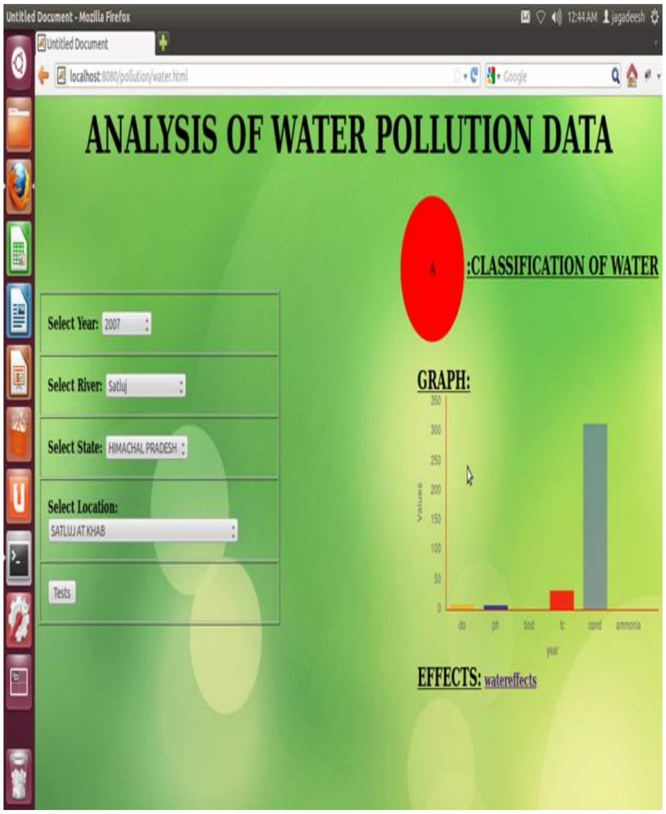
Typhoid

Malaria

Jaundice

Disruption of food-chain

A.2.3 WATER DATA ANALYSIS


Untitled Document - Mozilla Firefox

localhost:8080/pollution/water.html

ANALYSIS OF WATER POLLUTION DATA

A :CLASSIFICATION OF WATER

Select Year: 2007

Select River: Satluj

Select State: HIMACHAL PRADESH

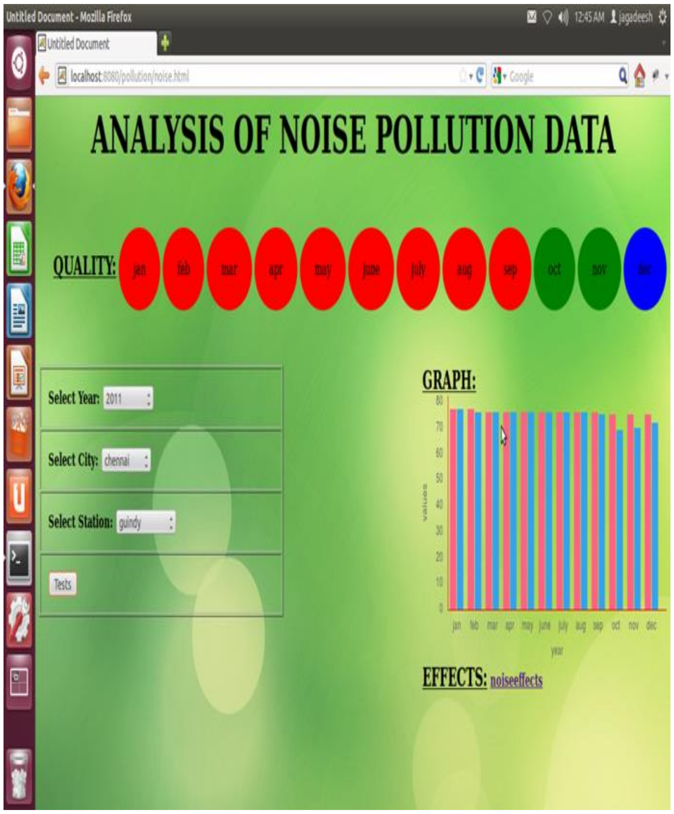
Select Location: SATLUJ AT KHAB

Tests

GRAPH:

EFFECTS: watereffects

A.2.4 NOISE DATA ANALYSIS


Untitled Document - Mozilla Firefox

localhost:8080/pollution/noise.html

ANALYSIS OF NOISE POLLUTION DATA

QUALITY: jan feb mar apr may june july aug sep oct nov dec

Select Year: 2011

Select City: chennai

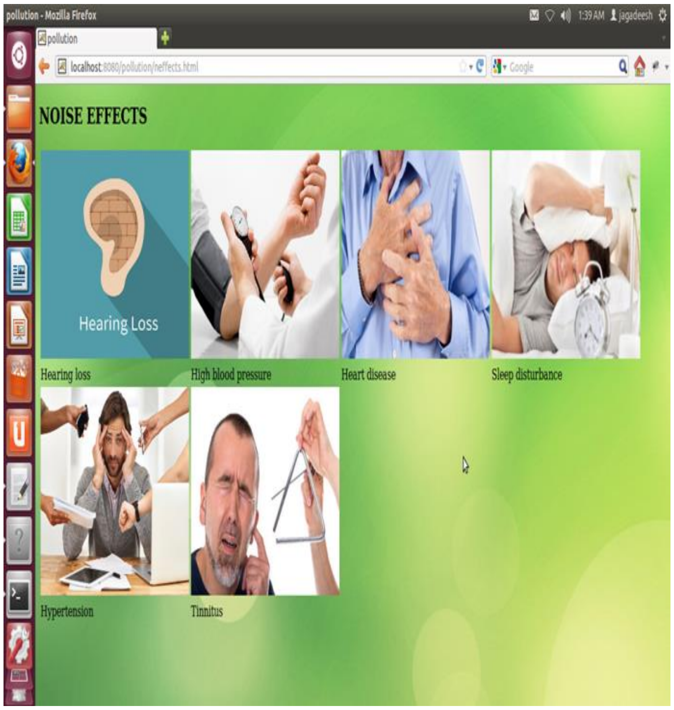
Select Station: guindy

Tests

GRAPH:

EFFECTS: noiseeffects

A.2.4.1 NOISE EFFECTS



REFERENCES

[1] M.Mazhar Rathore,Awais ahmad,Anand Paul,Seungmin Rho,IoT-based system for smart city development and urban planning using Big Data analytics,vol-2 ,Issue-5,2013.

[2] Nitin B Raut, Jabar H. Yousif, Sanad Al Maskari, and Dinesh Kumar Saini Cloud for Pollution Control and Global Warming ,WCE 2011, July 6 - 8, 2011, London, U.K .

[3] R.A.Roseline, Dr.P.Sumathi,Pollution Monitoring for Health Environment using Integrated Wireless Sensor Networks and Grid Computing,Volume 3, No. 1, January 2012.

[4] Sayed Nayab Basha,Akula Mallaiah,Embedded System to Control pollutants,vol-2 ,Issue-1,2012

[5] Z. Ghaemia, M. Farnaghib, A. Alimohammadib ,Hadoop Based Distributed System for Online Prediction of Air pollution based on Support Vector Machine,23–25 Nov 2015, Kish Island, Iran