**BIG DATA COURSEWORK 2**

**INVESTIGATION OF CORRELATION BETWEEN AIR POLLUTION AND RESPIRATORY DISEASE DATASETS USING HADOOP**

**AKINOSUN TEMITAYO A. (B00872232)**

**ARTIFICIAL INTELLIGENCE**

**SCHOOL OF COMPUTING**

**ULSTER UNIVERSITY**

**DECEMBER, 2022.**

**1.0 Introduction and Discussion of the Problem**

A World Health Organization report in 2008 asserts that 1.3 million global deaths were related to environmental air pollution; a figure which nearly tripled in 2012. Also, more than 2 million deaths have been attributed to air pollution annually [11]. The impact of air pollutants such as particulate matter, NO2, and SO2 have been reported to initialize and/or aggravate Chronic Obstructive Pulmonary Disease (COPD) among other respiratory diseases. The effects of COPD include increased mortality, reduction in pulmonary functionality, increase in infections, introduction of asthma, etc. [8]

In the United Kingdom, COPD is associated with the older and middle-aged population who smoke. Many of these groups do not even realize that they have the disease, making breathing problems worse over time, and thus more dangerous [2].

This report seeks to explore big data analysis tools on Hadoop environment with the aim of choosing the most suitable (with appropriate reasons) for the analysis of the selected datasets to find out the correlation between respiratory disease (COPD) and air pollution with focus on the United Kingdom.

**2.0 Background and Justification of Datasets Selection**

2.1 Air pollution

The dataset representing air pollution used for this coursework report was sourced from data.gov.uk. Though it includes information on mass substances release into controlled waters and sewers as well, the data manipulation and analysis on Hadoop has been used to choose only those released into air as required for the report. Also, 2018 data records were chosen in order to match the respiratory disease dataset chosen for the same year. This would help focus on 2018 particularly and derive any useful correlation between occurrence of Chronic Obstructive Pulmonary Disease (the respiratory disease considered) and air pollution with quantity of regulated substances released from industrial sites in the United Kingdom for that year.

2.2 Respiratory Disease

Chronic Obstructive Pulmonary Disease (COPD) has been mentioned as one of the most common diseases caused by air pollution [2]. It is a group of diseases including emphysema (damage to air sacs in the lungs) and bronchitis (long term inflammation of the airways), that causes difficulty in breathing due to damage to the lungs by air pollutants [5]. A dataset on drug prescription by the National Health Service (NHS) in December 2018 was used to analyze the frequency of COPD through the prescription of drugs used in managing it during the period under review. It gives the total amount of drugs prescribed and dispensed, net ingredient cost, actual cost, and total quantity. The prescriptions, identified by their British National Formulary (BNF) codes, were all written but some dispensed in England. Others were dispensed outside England but within the United Kingdom [6].

**3.0 Documentation of the Technical solution**

**3.1.1 Apache Zeppelin**

The tool that has been chosen for the analysis of this report is Apache Zeppelin. This web-based notebook supports interactive processes of data analysis including data loading, manipulation, sharing, and visualization on Spark and Hadoop. It supports programming languages like Python, Scala, HiveQL, Spark SQL and shell [1]. All the qualities mentioned above made it suitable for loading, joining, analyzing, visualizing and drawing insight from both 2018 air pollution dataset and COPD as the subject on respiratory disease. In addition, Zeppelin incorporates automatic SQL Context and Spark Context injection, and loading from maven library or local files is dependent on runtime jar.

Furthermore, operational functionalities like updating, cancelling and real-time display of ongoing work made Apache Zeppelin preferable for this coursework. Some basic statistical visualization like tables, charts, and plot are already on Zeppelin, thus visualization is gotten even without a Spark SQL query. Also, backend language outputs can be easily seen.

**Figure 1.** Control flow diagram illustrating Zeppelin Architecture [10]

Diagram

Description automatically generated

**3.1.2 Apache Hive**

The framework of the query language of hive, which is very much like SQL and runs without hitches on Hadoop made it gain its popularity. MapReduce complexity has also been addressed in this application and it can handle big data analysis on HDFS. The functions of hive include data summarization, data query and data analysis [9]. Tools for ease of data extraction, transformation and loading (ETL) are also provided on Hive. Apache utilizes a fast Optimized Row Columnar (ORC) storage system for Hadoop projects. Since it produces better performance to write, read and process data, ORC file format is a way in which Hive data can be efficiently stored [7].

**3.2 Other Available Technologies**

3.2.1 Hadoop Map Reduce

Although this big data processing infrastructure has been utilized over the years, the introduction of new technologies like Spark made its limitations obvious. It is slower in data analytics due to its two-stage paradigm, and does not adapt well for smaller data that can fit into a machine’s RAM. It is however more cost effective for processing a large amount of data.

3.2.2 Hadoop HDFS

Automatic failover is provided in the Hadoop File Distribution System (HDFS). Thus, in the event of hardware failure, the developer can retrieve data from another node in which the data also exists and across several programming languages thus ensuring data reliability. Since it stores data in multiple nodes, accessing and updating data is faster. It is also highly scalable, easy to set up and robust in implementation. However, the cost incurred during downtime can be on the high side and storing data in one location (even on different machines) increases the risk of data hacking [4].

3.2.3 Spark

Spark typically pulls out massive quantity data from Hadoop based data store to perform complex, in-memory data analysis in parallel in a short period of time. It runs faster compared to others because the input-output disk space usage and network processes have been reduced. It also supports different programming languages like Python, Java, and Scala in developing data analysis structures. The limitations of Spark stem from not having its own File Distribution System, high RAM requirement due to its in-memory operations, and no code optimization [4] .

**4.0 Analysis of the Datasets**

Below is a step-by-step description of the

**Prescription data, area addresses and air pollution data from data.gov.uk in 2018 were uploaded into Zeppelin using the command:**

*wget - -output-document project/<file name url>*

**Sample command to view the first 10 rows of data:**

*cd project*

*echo “PRESCRIPTION DATA \*\*\*\*\*\*\*\*\*\*”*

*cat gp\_prescription\_data\_2018\_11.csv | head -10*

**Joining data for both November and December of the same year:**

*cat project/gp\_prescription\_2018\_11.csv >> project/gp\_surgery\_address\_final.csv*

*cat project/gp\_prescription\_2018\_12.csv >> project/gp\_surgery\_address\_final.csv*

**Command transferring data to Hadoop HDFS:**

*hdfs dfs -mkdir project*

*hdfs dfs -put project/gp\_prescription\_data\_final.csv project/*

*hdfs dfs -put project/gp\_surgery\_address\_final.csv project/*

*hdfs dfs -put project/postcode\_region.csv project/*

*hdfs dfs -put project/pollution\_data.csv project/*

*hdfs dfs -put project/pollution\_site\_location.csv project/*

*hdfs dfs -ls project*

**Then, from Tabby (as root), I ran the following commands to enable jdbc (hive) to analyze and visualize the data on tables:**

*su hdfs*

*hdfs dfs -chown -R hive:hive /user/zeppelin/project*

**Creating a table:**

*%jdbc(hive)*

*DROP TABLE IF exists GP\_PRESCRIPTION\_COPD;*

*CREATE TABLE GP\_PRESCRIPTIONS\_COPD STORED AS ORC AS*

*SELECT bnf\_name, items, practice, period*

*FROM GP\_PRESCRIPTION\_DATA WHERE bnf\_name IN (‘Carbocisteine\_Cap 375mg’, ‘Carbocisteine\_Oral Soln 250mg/5ml’);*

**Loading the data:**

*%jdbc(hive)*

*LOAD DATA INPATH ‘/user/zeppelin/project/postcode\_region.csv’ into TABLE UK\_POSTCODE\_REGION;*

**Sample visualization code for the data:**

*%jdbc(hive)*

*SELECT \* FROM GP\_PRESCRIPTIONS\_COPD;*

The same codes as above were then used to load the other datasets: air pollution data, respiratory disease (COPD) dataset and including the postcode addresses and GP surgery addresses.

**5.0 Insights from Data Analysis**

It can be seen from the air pollution by postcode and region scatter diagram below (Figure 2) that the count in Lancaster (LA), North western England (CA), and Isle of Man (IM) is higher suggesting more industrial (manufacturing) activities or other causes of air pollution there.

Also, as shown in the pie chart below (Figure 3), the prescription of Carbocisteine, the drug for managing COPD is highest in the north west region. This confirms the earlier mentioned statement of higher air pollution in the area.

**Figure 2. Air Pollution count by Postcode Regions**

Chart

Description automatically generated

**Figure 3. COPD drug prescription by regions**

Graphical user interface, text, application, email, Excel

Description automatically generated

Another insight that is worthy of note is the bigger amount of prescription of Carbocisteine tablets (for adults) to its oral form (meant for children) as seen in the pie chart below (Figure 4). This is likely to be related to smoking habits amongst adults, and that adults frequent more at manufacturing sites than children.

Graphical user interface, application

Description automatically generated

**6.0 Conclusion and Discussion**

For the purpose of this big data analysis, Chronic Obstructive Pulmonary Disease was investigated as a leading respiratory disease and air pollution data location was the United Kingdom in the year 2018. From the analysis results and the insights drawn, it is like that there is a correlation between air pollution and the rate of respiratory diseases prevalent in the United Kingdom, especially from the respiratory disease related drugs being prescribed there. Analysis has been limited by access to credible and appropriately big data. Going forward, better grip on handling data on Hadoop and access to more appropriate data can help get more insight into this very important topic for the good of humanity.

**References**

[1] “Apache Zeppelin | cloudera,” *Cloudera*, 2022. [Online]. Available: https://www.cloudera.com/products/open-source/apache-hadoop/apache-zeppelin.html. [Accessed: 08-Dec-2022].

[2] “Chronic Obstructive Pulmonary Disease (COPD),” *NHS choices*, Sep-2019. [Online]. Available: https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/. [Accessed: 07-Dec-2022].

[3] “Chronic respiratory diseases,” *World Health Organization*. [Online]. Available: https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab\_1. [Accessed: 07-Dec-2022].

[4] “Data Warehouse Architecture - detailed explanation,” *InterviewBit*, 24-Jun-2022. [Online]. Available: https://www.interviewbit.com/blog/data-warehouse-architecture/. [Accessed: 07-Dec-2022].

[5] “Diseases caused by air pollution - risk factors and control methods,” *Air Filters for Clean Air*, 09-Feb-2018. [Online]. Available: https://cleanair.camfil.us/2018/02/09/diseases-caused-by-air-pollution-risk-factors-and-control-methods/. [Accessed: 07-Dec-2022].

[6] *GP practice prescribing data - presentation level*, 05-Mar-2019. [Online]. Available: https://www.data.gov.uk/dataset/176ae264-2484-4afe-a297-d51798eb8228/gp-practice-prescribing-data-presentation-level. [Accessed: 07-Dec-2022].

[7] J. Rafferty, “Practical 19: Hadoop HDFS + Hive,” in *Big Data and Infrastructure*, 02-Dec-2022.

[8] J. Samet, S. Buist, R. Bascom, J. Garcia, M. Lipsett, J. Mauderly, D. Mannino, C. Rand, I. Romieu, M. Utell, G. Wagner, D. V. Bates, M. L. Billingsley, M. Gelobter, B. F. Hobbs, S. Kleenberger, N. K. Leidy, S. London, W. F. McDonneli, D. Schwela, and J. C. Wiley, “What constitutes an adverse health effect of air pollution?,” *Oregon Health & Science University*, 01-Jan-1970. [Online]. Available: https://ohsu.pure.elsevier.com/en/publications/what-constitutes-an-adverse-health-effect-of-air-pollution-2. [Accessed: 07-Dec-2022].

[9] R. Sharma, “Hive vs spark: Difference between hive & spark [2021] | upgrad blog,” *UpGrade*, Nov-2022. [Online]. Available: https://staging-mp.upgrad.dev/blog/hive-vs-spark/. [Accessed: Dec-2022].

[10] T. A. S. Foundation, “Submarine interpreter for Apache Zeppelin,” *Sitewide ATOM*. [Online]. Available: https://zeppelin.apache.org/docs/0.9.0/interpreter/submarine.html. [Accessed: 08-Dec-2022].

[11] X. Q. Jiang, X. D. Mei, and D. Feng, “Air pollution and chronic airway diseases: What should people know and do?,” *Journal of thoracic disease*, Jan-2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26904251/. [Accessed: 07-Dec-2022].