AIR QUALITY INDEX AND RESPIRATORY DISORDER ANALYSIS USING HADOOP-HIVE

Athulya Shaji

MSc Computer Science

School of Computing, Engineering, and
Built Environment

Ulster University

Belfast

Shaji-A@ulster.ac.uk

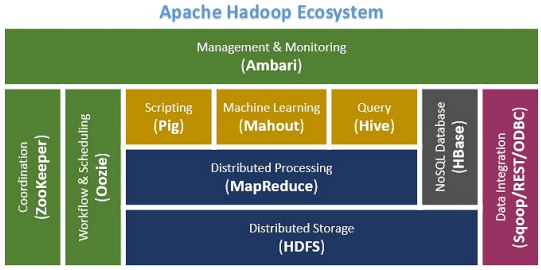
Abstract— **This paper focuses on the selection, cleaning, processing, and analysis of large amounts of data associated with air pollution and respiratory disorders to arrive at a meaningful insight. The process of handling this huge data aka Big Data is done with the help of various software tools and open-source resources. The software used here is Hadoop for the processing of this massive amount of data. Wherein Hive is used for the cleaning and processing and Zeppelin is used for the visualization of data as they both have an SQL-like query system to handle large nonrelational datasets.**

Keywords—Dataset, SQL, Hadoop, Hive, Zeppelin, Query, Visualization.

I. INTRODUCTION

The rate of data being collected every second is ever-increasing and is the highest of all time. The data coming in from various devices and sensors in this era of the internet of things is enormous. The storage, processing, and analyzing of this data to draw meaningful full conclusions are of great challenge. These data may be structured, semi-structured, or unstructured. They possess the characteristics of high velocity, volume, and variety. This data is nothing but popularly known as big data.

The software used for handling a large amount of data in this project is Hadoop. Hadoop is an open-source software utility that uses a distributed file system (HDFS) to store data in large clusters of systems. The retrieving and processing of data is done using the MapReduce method. To use Hadoop, we do not have to be part of this large cluster, rather with the help of the Docker container Hadoop can the used from any kind of system. The user-friendly web UI Ambari makes it so much easier to work on Hadoop services.



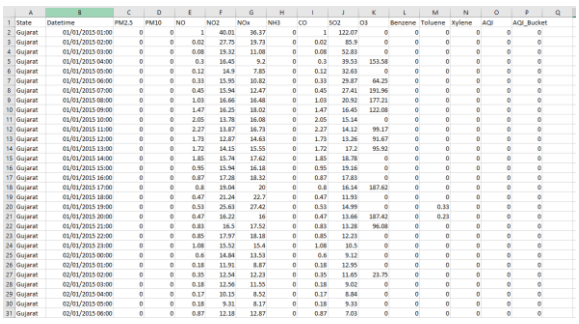
Hadoop Architecture [1]

Hive is an SQL-like query-based interface to read, write, and process the data stored in HDFS. The biggest advantage of the hive is the SQL-like querying itself which makes it the most flexible software when dealing with large structured data like the one used for this project. The visualization tool Zeppelin again provided by Hadoop is a wonderful software to visualize data in different ways without actually writing long code to get the same results, these pictorial representation gives a better insight into the data.

II. PROPOSED SYSTEM

A. Selected Datasets

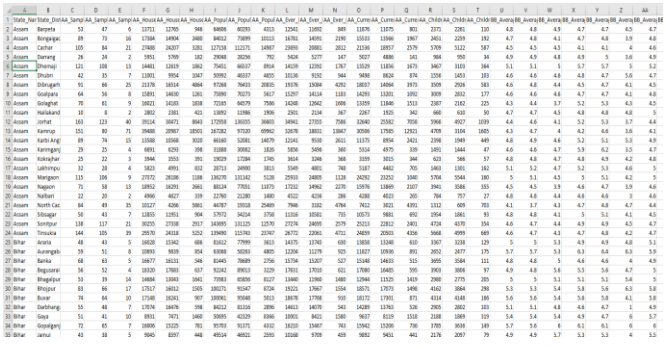
This project aims to process data regarding air pollution and analyze the major pollutants involved in the determination of the air quality index of a particular region and how this becomes a crucial part of the cause of various health hazards, especially respiratory disorders. There are two main datasets used in this project, both taken from Kaggle (dataset repository). The first dataset consists of the readings of different air pollutants such as PM2.5, PM10, SO, NOX, NO2, etc, and a final air quality index value of different states in India for every hour from the year 2015 to 2020. This is a very large dataset consisting of more than seventy thousand rows.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	State	Datetime	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket	
2	Gujarat	01/01/2015 01:00	0	0	1	40.01	36.37	0	1	122.07	0	0	0	0	0	0	
3	Gujarat	01/01/2015 02:00	0	0	0.02	27.75	19.73	0	0.02	85.9	0	0	0	0	0	0	
4	Gujarat	01/01/2015 03:00	0	0	0.08	19.32	11.08	0	0.08	52.83	0	0	0	0	0	0	
5	Gujarat	01/01/2015 04:00	0	0	0.3	16.45	9.2	0	0.3	39.53	153.58	0	0	0	0	0	
6	Gujarat	01/01/2015 05:00	0	0	0.12	14.9	7.85	0	0.12	32.63	0	0	0	0	0	0	
7	Gujarat	01/01/2015 06:00	0	0	0.33	15.95	10.82	0	0.33	29.87	64.25	0	0	0	0	0	
8	Gujarat	01/01/2015 07:00	0	0	0.45	15.94	12.47	0	0.45	27.41	191.96	0	0	0	0	0	
9	Gujarat	01/01/2015 08:00	0	0	1.03	16.66	16.48	0	1.03	20.92	177.21	0	0	0	0	0	
10	Gujarat	01/01/2015 09:00	0	0	1.47	16.25	18.02	0	1.47	16.45	122.08	0	0	0	0	0	
11	Gujarat	01/01/2015 10:00	0	0	2.05	13.78	16.08	0	2.05	15.14	0	0	0	0	0	0	
12	Gujarat	01/01/2015 11:00	0	0	2.27	13.87	16.73	0	2.27	14.12	99.17	0	0	0	0	0	
13	Gujarat	01/01/2015 12:00	0	0	1.71	12.87	14.63	0	1.71	13.26	91.67	0	0	0	0	0	
14	Gujarat	01/01/2015 13:00	0	0	1.72	14.15	15.55	0	1.72	17.2	95.92	0	0	0	0	0	
15	Gujarat	01/01/2015 14:00	0	0	1.85	15.74	17.62	0	1.85	18.78	0	0	0	0	0	0	
16	Gujarat	01/01/2015 15:00	0	0	0.95	15.94	16.18	0	0.95	19.16	0	0	0	0	0	0	
17	Gujarat	01/01/2015 16:00	0	0	0.87	17.28	18.32	0	0.87	17.83	0	0	0	0	0	0	
18	Gujarat	01/01/2015 17:00	0	0	0.8	19.04	20	0	0.8	16.14	187.62	0	0	0	0	0	
19	Gujarat	01/01/2015 18:00	0	0	0.47	21.24	22.7	0	0.47	11.91	0	0	0	0	0	0	
20	Gujarat	01/01/2015 19:00	0	0	0.53	25.63	27.42	0	0.53	14.99	0	0	0.33	0	0	0	
21	Gujarat	01/01/2015 20:00	0	0	0.47	16.22	16	0	0.47	13.66	187.42	0	0.23	0	0	0	
22	Gujarat	01/01/2015 21:00	0	0	0.83	16.5	17.52	0	0.83	13.28	96.08	0	0	0	0	0	
23	Gujarat	01/01/2015 22:00	0	0	0.85	17.97	18.18	0	0.85	12.23	0	0	0	0	0	0	
24	Gujarat	01/01/2015 23:00	0	0	1.08	15.52	15.4	0	1.08	10.5	0	0	0	0	0	0	
25	Gujarat	02/01/2015 00:00	0	0	0.6	14.84	13.53	0	0.6	9.12	0	0	0	0	0	0	
26	Gujarat	02/01/2015 01:00	0	0	0.18	11.91	8.87	0	0.18	12.95	0	0	0	0	0	0	
27	Gujarat	02/01/2015 02:00	0	0	0.35	12.54	12.23	0	0.35	11.65	23.75	0	0	0	0	0	
28	Gujarat	02/01/2015 03:00	0	0	0.18	12.56	11.55	0	0.18	9.02	0	0	0	0	0	0	
29	Gujarat	02/01/2015 04:00	0	0	0.17	10.15	8.52	0	0.17	8.84	0	0	0	0	0	0	
30	Gujarat	02/01/2015 05:00	0	0	0.18	9.31	8.17	0	0.18	9.33	0	0	0	0	0	0	
31	Gujarat	02/01/2015 06:00	0	0	0.87	12.18	12.87	0	0.87	7.03	0	0	0	0	0	0	

The second dataset is a large dataset consisting of many key social parameter indicators of different states of India during

the year 2018. The dataset consists of more than thirty columns, which consist of measurements of respiratory infection, asthma, and many other health disorders.



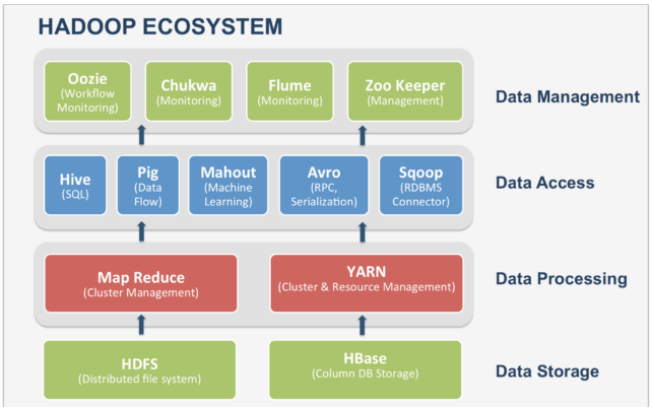
The two above-mentioned datasets are very much apt for further analysis in this project as both have the Indian states as the key column and the data collection time is overlapping.

III. SOLUTION PRODUCED

I have given a Hadoop Hive and Zeppelin-based solution for the analysis of the air pollution in India during the time period of 2015 – 2020 and the respiratory disorders in the Indian states.

Hadoop consist of three components:

1. Hadoop Distributed File System (HDFS) is the storage unit
2. MapReduce is the processing unit
3. YARN is the resource management unit



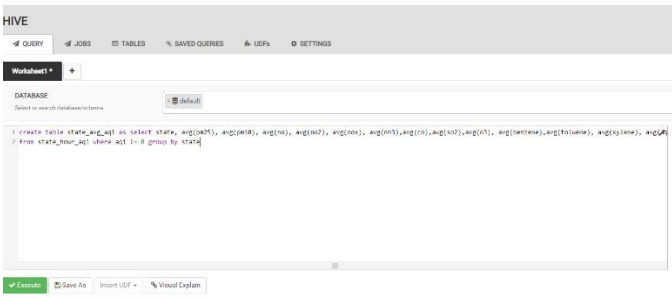
Hadoop components [2]

The data is visualized using the Zeppelin software, which is a web-based notebook for data exploration, visualization, and data sharing.

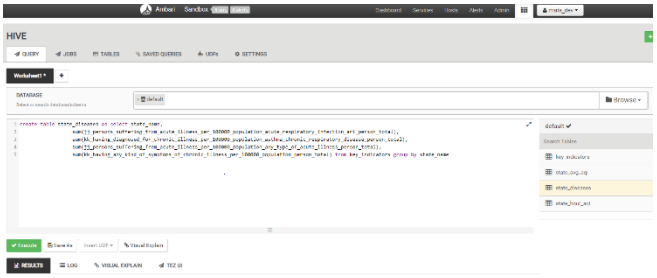
IV. ANALYSIS OF DATA

As the two main datasets used in this project have been already introduced in the previous section. I had to replace few null values in the dataset with zero inorder to avoid any errors while applying SQL functions for comparisons.

1. The air quality dataset was uploaded to Hive and modified to create a new table using the below query to bet a better-confined data by calculating the average of the pollutants by grouping w.r.t. states.



2. The key indicator dataset was also uploaded and a new table was derived from it with only the columns which are relevant to this analysis using the below query again by adding up all the respiratory cases reported in different states grouping by state

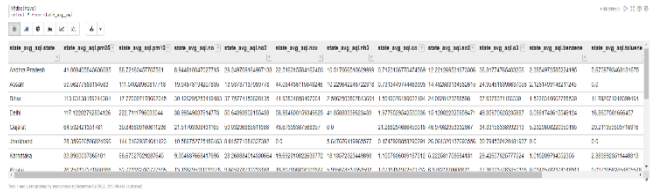


V. VISUALIZATION

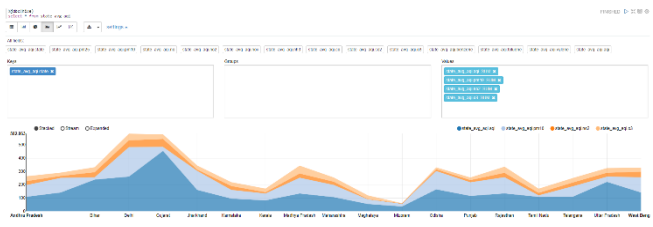
As discussed earlier using the zeppelin tool is very easy to explore the data and get meaningful insights.

The tables which were derived from the original csv files in Hive will be used further for visualization and interpretation.

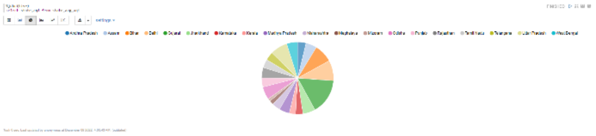
1. The data in the derived air quality table is been displayed below.



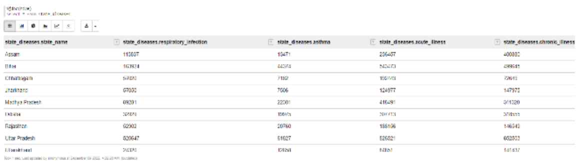
2. The area chart upon trying different combinations of values gives a conclusion that PM10 is the most significant pollutant for the AQI in the Indian stated according to the dataset.



3. Following is the pie chart diagram showing the AQI of the states from which Gujarat is the highest and Mizoram is the lowest

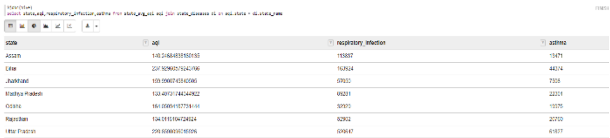


4. The data of derived from the key indicators showing the respiratory disorders



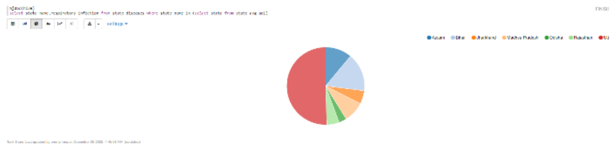
state_diseases.state_name	state_diseases.respiratory_infection	state_diseases.asthma	state_diseases.acute_illness	state_diseases.chronic_illness
Assam	113887	13471	236407	400880
Bihar	163924	44374	545373	498645
Chhattisgarh	[illegible]	7182	[illegible]	79649
Jharkhand	67030	7606	124977	147975
Madhya Pradesh	69201	22301	416491	311320
Odisha	[illegible]	19975	[illegible]	[illegible]
Rajasthan	62902	20760	188106	146549
Uttar Pradesh	520647	61827	826521	852303
Uttarakhand	[illegible]	[illegible]	[illegible]	141437

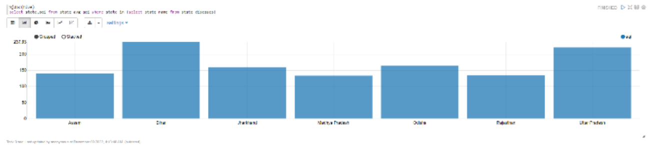
5. The below is the combination of the data from both the table to understand the relation of AQI on respiratory disorders.

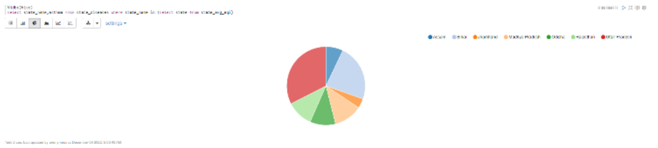


state	aqi	respiratory_infection	asthma
Assam	140.24684838150193	113887	13471
Bihar	237.92960579243706	163924	44374
Jharkhand	159.9900745840506	67030	7306
Madhya Pradesh	133.40731744244922	69201	22301
Odisha	164.05094187731444	32979	19975
Rajasthan	134.0115104724524	62902	20760
Uttar Pradesh	220.8598036015006	520647	61827

6. The following pie charts and bar diagram shows the direct correlation between AQI and respiratory infection and asthma.







Conclusion and Insight

In this paper, I have discussed about the growing intensity of data and one of the many methodologies to store, process, visualize and analyze big data using Hadoop.

From the solution implemented in this paper, I was very much elevated with the conclusion that would be drawn by the end. It was very clear that the AQI of a place is closely affecting the health of the people living in that particular place. This insight was driven by two different large sets of datasets using simple queries and visualization

Reference

[1] Gumlet.Io. Retrieved 9 December 2022, from https://dezyre.gumlet.io/files.dezyre.com/images/blog/Big+Data+and+Hadoop+Training+Hadoop+Components+and+Architecture_1.png?w=640&dpr=1.3Apache Hadoop. (n.d.). Apache.Org. Retrieved January 12, 2022, from https://hadoop.apache.org

[2] Perficient.com. Retrieved 9 December 2022, from https://blogs.perficient.com/files/Hadoonh-3-