Susceptibility Analysis of Novel Corona Virus Using Hadoop Distributed File System

Pravin Kumar Samanta, Soumyadev Mukherjee, and Nirmal Kumar Rout

Abstract  The Novel Corona Virus, COVID-19 is a major setback to the progress of the nation. The year 2020 has witnessed a disappointing beginning with this deadly virus having taken away many innocent lives. The COVID-19, which happens because of the virus Corona, showcases symptoms which are mild and at times doesn’t even reflect any. However, a small section falls severely ill, and some eventually perish. This proposed research work is implemented to identify who is having the highest risk of serious illness and who are more susceptible to this disease. In this proposed research, a huge amount of semi structured healthcare data is stored and processed which are collected from healthcare dataset. Hence, in the proposed work ‘Hadoop’ is utilized for processing the data gathered. The input data is processed using MapReduce and finally the result is loaded into the Hadoop Distributed File System (HDFS). After details analysis and justification with the help of ‘Hadoop’ proposed work has predicted many possible susceptible cases which can help to reduce the number of fatalities and also save human lives.

Keywords  Healthcare · Hadoop · MapReduce · HDFS · Tableau · Big data · COVID-19 · Corona virus
1 Introduction

The Novel Corona virus, COVID-19 is an immensely infectious disease resulting from a very recently found Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) [1]. First originated in Wuhan, China in late December, 2019 it has now been characterized as a global pandemic [2]. An individual infected with the COVID-19 virus generally witnesses mild to slight respiratory issues and get well without the need of any additional medical care. The aged mass, and individuals possessing hidden health issues like heart problems, diabetes, acute respiratory illness, and cancer are more highly prone to fall severely ill [3]. Now the question may arise that who is at most risk for COVID-19? It is high time for the nation to figure out who is more prone to fall prey to this deadly virus. Preventive measures should be taken and they should be handled with intensive care to reduce the number of fatalities.

With reference to an analysis of 45,000 positive cases in China, less than 1.0% of physically fit individuals who fell prey to the COVID-19 perished as a consequence of the same. However, it is close to 6.0% for patients having cancer, high pressure or acute respiratory illness, 7.3% for people with high glucose content and 10.5% for heart patients [1]. Patients aged 80 or more are at high risk, with close to 14.8% expiring. Younger children, those aged below 9, are not prone to serious illness. Another research suggested that males are more susceptible to severe illness than females, but those are hypothetical and may showcase the fact that males smoke more than females in China [4–6]. Coming to India, the primary concern is to analyse the trend and take intensive care of the people who are more likely to fall prey to the Novel Corona Virus. The sole objective of this proposed work is identifying the one who is most susceptible and to take preventive measures to protect them using the help of healthcare dataset. This research paper follows the following pattern: Sect. 2 describes the Hadoop Distributed File System (HDFS) methodology. Related work in the application of healthcare is described in Sects. 3 and 4 presents the description of the proposed model. Section 5 describes the implementation and results; Sect. 6 offers the conclusion of the work with future scope.

2 HDFS Methodology

Hadoop offers five daemons with each daemon possessing a Java Virtual Machine—(A) DataNode, (B) NameNode, (C) JobTracker, (D) Secondary NameNode, (E) TaskTracker. Demons which store data and metadata, i.e.; DataNode and NameNode, come under a part of Hadoop Distributed File System (HDFS). The TaskTracker and JobTracker, which keep track and actually execute the job, come under MapReduce layer. The HDFS is used in this proposed research work because of the following reasons [7].
**Large Dataset:** Current population of India is more than 1.3 Billion, if we want to analyze data for that amount of people then the dataset will be huge. That much huge amount of data can’t be processed by normal file system. That’s why to get a smooth workflow the HDFS is used for analysis [8].

**Data Replication:** For working with a large dataset, occurrences of unfortunate situations like hardware failure, crashing of a node are pretty common. In such situations data loss is occurred. To overcome this kind of problem HDFS is providing a feature called data replication. The data is copied across numerous nodes in the cluster by the creation of duplicates. This methodology is maintained across stipulated time intervals by HDFS and the duplication process is taken care of by the same. The moment as machine in the cluster crashes, the data should be retrieved from other machines. Loss of data is far sighted threat and almost negligible.

**Scalability:** Our main goal of the work is to analyze healthcare dataset using Hadoop and facilitate a smoother conduct of the fight against COVID-19. So, the proposed work is scalable in order make it a dynamic project. This is achieved by using HDFS. In HDFS the infrastructure is scaled up by adding more racks or clusters to this system.

**Data Locality:** In older systems, the data is brought at the application layer and then worked upon. In this proposed research work, as a consequence of the huge bulk of data, bringing data to the application layer has lowered down the overall performance.

In HDFS, the computation part is brought to the Data Nodes where data resides. Hence, with Hadoop HDFS, computation logic is not moved to the data, rather than data is moved to the computation logic.

### 3 Related Work

The sole aim of the ‘Healthcare System’ is the treatment, diagnosis of health-related problems of human beings. Recently big data analytics is playing a very important role for the development of current healthcare practices and research [9]. For the processing of huge amount of healthcare data record lots of efficient tools and methodologies are required. For the storage and processing of huge volumes of data set, Hadoop is playing a significant role for modern healthcare system [7]. Meena et al. proposed a system for the analysis of Healthcare data set against various queries using Hadoop Cluster and Pig Latin, where in that method fully distributed Hadoop Cluster mode was utilized [10]. Ashwin Belle et al. designed a model to infer fluctuations in enrolment due to the effect of carelessness using Hadoop and Hive [11]. Highly based on Market Research Sabyasachi Dash et al. proposed a prototype for the retrieval of patient information using Sqoop (SQL + Hadoop) and Hive [12]. A. S. Thanuja Nidhadi et al. utilized data-mining techniques in classification of Naive Bayesian (NB), and Back Propagation (BP) to identify patterns leading to fraud and emphasized on using Hadoop HDFS solution for storage and processing of medical data to avoid modern issues in healthcare [13]. Based on statistical summaries and
data quality visuals Wullianallur Raghupathi et al. [14] suggested a proposal to use HDFS in data calibration, data wrangler in data cleansing, R in data analytics and MapReduce Tools in Architecture Testing. In this proposed method the HDFS is used for identifying the susceptibility analysis of Corona virus based on Healthcare data set.

4 Description of Proposed Model

About the dataset:
In this proposed method Healthcare data is used which is collected from publicly published dataset [15]. The dataset comprises of factual information, results of examinations and information given by the patients. ID refers to the patient identification. Age, height, weight and gender are objective features which are in days, cm, kg and categorical code respectively. Some of the subjective features like asthma count and smoking reflect binary outputs, cholesterol count which is an examination feature showcases three different scenarios wherein (1) refers to normal cholesterol, (2) refers to high cholesterol and (3) refers to way above normal cholesterol. Corona count reflects the chances of the patient being tested positive for COVID-19 provided he’s suffering from one of the diseases.

The dataset is divided into smaller blocks which are primarily processed by “Map Phase” in parallel and then by “Reduce Phase”. Hadoop framework has sorted out the output of the Map phase which are then given as an input to Reduce Phase to initiate parallel reduce tasks. These input and output files are stored in file system (Fig. 1).

Input dataset is adapted by MapReduce framework from HDFS. Input dataset is taken as key-value pair and broken down for effective analysis. Then, the number of positive cases is mapped to corresponding diseases and shuffled accordingly. The number of positive cases is the reduced output. The final result is sent to the authorities and hospitals for preventive measures. In this work five MapReduce key functions are used to get the desired output/outcome. The MapReduce job model is described in Fig. 2. The functionalities of each functions are as follows:

![Fig. 1 Sample of input healthcare dataset.](https://www.kaggle.com/iamhungundji/covid19-symptoms-checker)
Fig. 2  Map reduce job model description (adopted from [10])

- **UID_Mapper.java**: Filters the header and writes to mapper output.
- **UID_Reducer.java**: Aggregate values for each disease and outputs positive cases.
- **Sort_Mapper.java**: Receive output from previous UID_Mapper-Reducer phase and shuffle \((x, y1)\) and \((x, y2)\) pair into \((x, (y1, y2))\) form.
- **Sort_Comparator.java**: Sorts and reduces the output in descending order.
- **Sort_Reducer.java**: Swap \((x, y)\) pair into \((disease, count)\) format and produce output.
- **Driver.java**: It’s the main driver program for the MapReducer job.

Use case diagrams of the proposed model (which is described in Fig. 3) outlines the usefulness of a framework utilizing the on-screen characters and use cases. The actors here are the ‘Users’ and the ‘Admin’ of the system. Login, import dataset, display data, train dataset etc. are the various use cases.

## 5 Implementation and Results

Having done intensive analysis, the following results are found out which correspond to the dataset (which has already been discussed) anatomization using Hadoop.

Detailed description of the outcomes from the implementation results are given in Figs. 5, 6, 7, 8 and 9. In Fig. 4 gender wise height to weight distribution is represented. In Fig. 5 to Fig. 9, the number of negative cases is mapped to binary ‘0’ as represented by blue color and the number of positive cases is mapped to binary ‘1’ as represented by orange color (corona virus negative signifies as ‘0’, and positive as ‘1’). Figure 5 shows the gender wise count of COVID-19 cases. The graph plotted reflects the fact that individuals belonging to gender 2 (male) over gender 1 (female) are more prone to fall prey to COVID-19. In Fig. 6 the cholesterol versus count of COVID-19 cases are presented where ‘type-1’ refers to normal cholesterol, ‘type-2’ refers
to high cholesterol and ‘type-3’ refers to way above normal cholesterol. The graph plotted clearly depicts the fact that people with above normal and way above normal cholesterol are highly susceptible to the dangerous COVID-19. Figure 7 presents the Glucose content of normal (type-1), high (type-2) and very high (type-3) versus count of COVID-19 cases. The graph depicts that individuals with high or very high glucose content are more likely to fall prey to dangerous COVID-19. Figure 8, the graph of smokers versus count of COVID-19 cases gives a clear indication that (type 0) smokers or smokers are at a higher risk than non-smokers (type-1). In the plot of Fig. 9 asthma patients versus count of COVID-19 cases signifies that type-0 asthma patients are at a severe risk from COVID-19.

6 Conclusion and Future Scope

In this proposed method, after implementation of the model using Hadoop, HDFS and detailed analysis of the Healthcare dataset, it is observed that males over females and people with high glucose, cholesterol, asthma patients, smokers etc. are highly
susceptible to the COVID-19. This proposed research has the ability to segregate out patients with high or very high glucose content, smokers, asthma patients, people with above normal cholesterol and pay primary attention to them as they’re at a higher risk of severe illness. This method has the ability to help the Government, authorities and healthcare units significantly to figure out the associated phenomenon which adds to the intensity of the dangerous COVID-19 to reduce the number of fatalities. This is also very much helpful for the healthcare units to make amendments to the existing database, to take precautionary measures and ensure a better tomorrow. The nation demands a better tomorrow and this proposed approach is a small step to bring
Fig. 6  Cholesterol versus count of COVID-19 case for type-1, type-2 and type-3

Fig. 7  Glucose versus count of COVID-19 cases for type-1, type-2 and type-3

back the lost glory. Let the innocent faces smile again and let us all say, “We shall overcome someday!”
Fig. 8  Smokers versus count of COVID-19 cases

Fig. 9  Asthma patients versus count of COVID-19 cases

References

1. Zhao, S., Lin Q., Ran, J., et al.: Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. Int. J. Infect. Dis., (2020)
2. Liangjun, Chen., Weiyong, Liu., Qi, Zhang., et al.: RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg. Microbes. Infect. 9, 313–319 (2020)
3. Zu, Z.Y., Jiang, M.D., Xu, P.P., Chen, W., Ni, Q.Q., Lu, G.M., Zhang, L.J.: Coronavirus disease 2019 (COVID-19): a perspective from china. Radiology, (2020)
4. https://radiopaedia.org/articles/covid-19-3 and https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200413-sitrep-84-covid-19.pdf?sfvrsn=44f511ab_2
5. Liang, T.: Handbook of COVID-19 Prevention and Treatment (2020)
6. Luccioni, A., Bullock, J., Pham, K.H., Lam, C., Luengo-Oroz, M.: Mapping the landscape of artificial intelligence applications against COVID-19. arXiv:2003.11336 (2020)
7. Qin Yao, Yu., Tian, Peng-Fei Li., Tian, Li-Li., Qian, Yang-Ming, Li, Jing-Song: Design and development of a medical big data processing system based on hadoop. J. Med. Syst. 39(3), 23 (2015)
8. Bizer, Christian., Boncz, Peter., Brodie, Michael L., Erling, Orri: The meaningful use of big data: four perspectives four challenges. ACM Sigmod Rec. 40(4), 56–60 (2012)
9. Senthilkumar, S.A., Rai, B.K. Meshram, A.A., Gunasekaran, A., Chandrakumarmangalam, S.: Big data in healthcare management: a review of literature. Am. J. Theor. Appl. Bus. 4(2), 57–69 (2018)
10. Meena, I., Aravind, R., Sarker, V., Sanjudharan, M.S.: Healthcare analysis using hadoop framework. IJSART4(14) 300–306 (2018)
11. Belle, A., Thiagarajan, R., Soroushmehr, S.M., Navidi, F., Beard, D., Najarian, K.: Big data analytics in healthcare. BioMed Res. Int. (2015)
12. Dash, S., Shakyawar, S., Sharma, M., Kaushik, S.: Big data in healthcare: management, analysis and future prospects. J. Big Data 6 (2019)
13. Thanuja Nishadi, A.S.: Healthcare big data analysis using hadoop MapReduce. Int. J. Sci. Res. Pub. 9(3), 87104 (2019). http://doi.org/10.29322/IJSRP
14. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: proise and potential. Health Inf. Sci. Syst. 2(3), 7 (2014). https://doi.org/10.1186/2047-2501-2-3
15. https://www.researchgate.net/publication/333201013_Healthcare_Analysis_Using_HadoopFramework