Big Data Analysis for Predictive Healthcare Information System

Neha Maurya[1], Pankaj Pratap Singh[2], Amit Kishor[3], Anirudh Kumar Tripathi[4]
[1]*Departmentof Computer Science and Engineering, Swami Vivekanand Subharti University*
[234]*Departmentof Computer Science and Engineering, Swami Vivekanand Subharti University*

Abstract- This paper surveys big data with highlighting the big data analytics in medicine and healthcare. Healthcare big data refers to the vast quantities of data that is now available to healthcare providers. As a response to the digitization of healthcare information and the rise of value-based care, the industry has taken advantage of big data and analytics to make strategic business decisions. Faced with the challenges of healthcare data volume, velocity, variety, and veracity, health systems need to adopt technology capable of collecting, storing, and analysing this information to produce actionable insight.

Keywords- Big Data, Android, Python, Hive, Django, Hadoop, Big Data Mining, Predictive Analytics.

I. INTRODUCTION

Big Data has changed the way we manage, analyse and leverage data in any industry. One of the most promising areas where it can be applied to make a change is healthcare. Healthcare analytics have the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general. Average human lifespan is increasing along world population, which poses new challenges to today's treatment delivery methods. Health professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and look for best strategies to use these numbers. In this paper, we would like to address the need of big data in healthcare

II. What Is Big Data in Healthcare?

The application of big data analytics in healthcare has a lot of positive and also life-saving outcomes. Big data refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analysed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc.

Now that we live longer, treatment models have changed and many of these changes are namely driven by data. Doctors want to understand as much as they can about a patient and as early in their life as possible, to pick up warning signs of serious illness as they arise – treating any disease at an early stage is far simpler and less expensive. With healthcare data analytics, prevention is better than cure and managing to draw a comprehensive picture of a patient will let insurances provide a tailored package. This is the industry's attempt to tackle the siloes problems a patient's data has: everywhere are collected bits and bytes of it and archived in hospitals, clinics, surgeries, etc., with the impossibility to communicate properly.

Indeed, for years gathering huge amounts of data for medical use has been costly and time-consuming. With today's always-improving technologies, it becomes easier not only to collect such data but also to convert it into relevant critical insights, that can then be used to provide better care. This is the purpose of healthcare data analytics: using data-driven findings to predict and solve a problem before it is too late, but also assess methods and treatments faster, keep better track of inventory, involve patients more in their own health and empower them with the tools to do so.

III. BRIEF DESCRIPTION ABOUT THE RESEARCH:

- There are lots of diseases that occurs every season and are common so with the help of different datasets I am going to analyse what are the chances of any disease to occur in future and going to aware people about it.
- As we all know that due to unpredictive increase in occurrence of diseases availability of doctors get decreases so I am planning to aware people about it and making it easy to cope such kind of problems.
- A feeling bar will be there in order to get information about how an individual is feeling.
- Also providing the list of basic medicines related to symptoms of diseases in order to rectify the diseases

Hadoop -

Hadoop is an open source framework from Apache and is used to store process and analyse data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover, it can be scaled up just by adding nodes in the cluster.

Modules of Hadoop

- **HDFS:** Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
- **Yarn:** Yet another Resource Negotiator is used for job scheduling and manage the cluster.

- **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
- **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

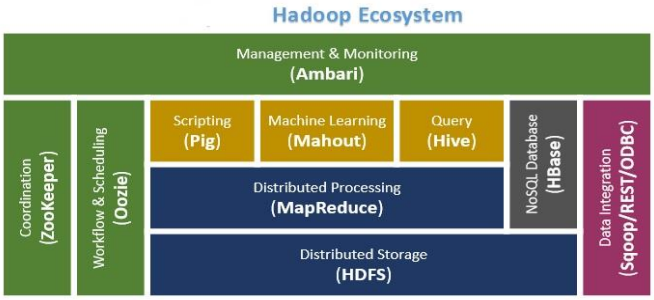


Fig.1: (Hadoop Ecosystem)

PYTHON

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. Python source code is also available under the GNU General Public License (GPL).

Python can be used on a server to create web applications and software to create workflows. Python can connect to database systems. It can also read and modify files. It is used to handle big data and perform complex mathematics.

Features:

- **Easy to learn**
- **Easy to read**
- **Easy to maintain**
- **Portable**
- **GUI Programming**
- **A broad standard library**

Python can be easily integrated with C, C++, Hadoop, CORBA and Java. It supports functional and structured programming methods as well as OOP.

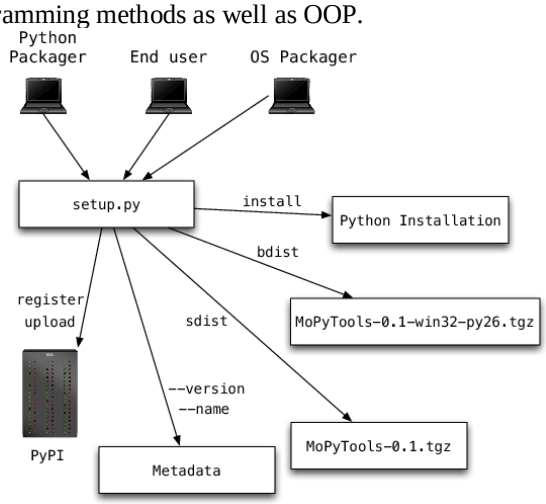


Fig.2: (Python)

PyHive

PyHiveis a python module which is used to run the HiveQL (HQL) over the hive through the python script.

Django

Django is a web development framework that assists in building and maintaining quality web applications. Django helps eliminate repetitive tasks making the development process an easy and time saving experience. This tutorial gives a complete understanding of Django. Django's primary goal is to ease the creation of complex database-driven websites.

Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Django makes it easier to build betterweb apps quickly and with less code.

Feature:

- Python web-framework
- High Scalability
- Versatile in Nature
- Offers High Security
- Provides Rapid Development

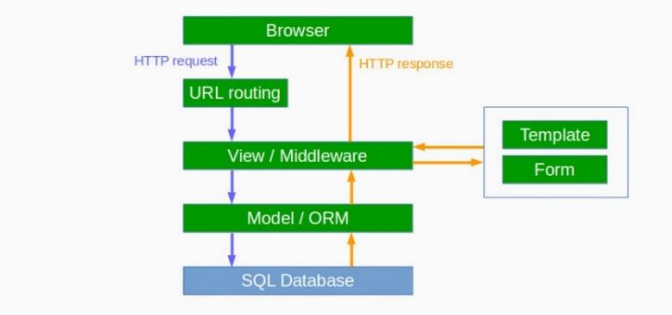


Fig.3: (Django Architecture)

Android

- **Definition and properties**

[7] Android is a mobile operating system (OS) first developed by a Silicon Valley company by the name of Android Inc. A collaboration spearheaded by Google in 2007 through the Open Handset Alliance (OHA) gave Android an edge in delivering a complete software set, which includes the main OS, middleware and specific mobile application, or app.

- **Android Features & Specifications**

Android is a powerful Operating System supporting a large number of applications in Smart Phones. These applications make life more comfortable and advanced for the users. Hardware that support Android are mainly based on ARM architecture platform.

- **User Interface & Navigation**

[9] Your app's user interface is everything that the user can see and interact with. Android provides a variety of pre-built UI components such as structured layout objects and UI controls that allow you to build the graphical user interface for your app. Android also provides other UI modules for special interfaces such as dialogs, notifications, and menus.

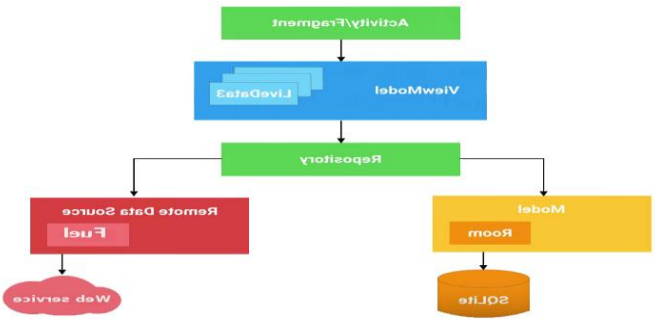


Fig.4:

Module 1: Mobile will automatically fetch date, time, location which is then utilized in identification of season. Once season is identified then disease analysis will take place using hive tool based on past disease data of that region which will result into list of diseases in a priority order of their occurrence. Then analysis of disease doctor availability take place using hive tool (which will be done on doctors' data of that region) that will result into doctors who will be available in that particular season (for curing disease) And simultaneously basic medicine for curing diseases will be identified from the medicine data of diseases.

Module 2: It is the manual section of my application in which user will enter their disease symptoms which will then be utilised for identification of disease from which the user is striving. This process will be done by analysing disease data for symptoms that are entered by user using hive tool and result will be the list of diseases with best possible match in a priority order then a kind of feedback will be taken from the user about their feelings. If they rate their feelings less than 5 then analysis will be done corresponding to that disease on the disease data using hive tool for the identification of basic medicines and user medical course will be started. After few days feedback will be taken again in words and if it is not good then analysis will be done on doctors' data for identification of doctors (of that particular disease) And in case of good feedback the medical course will be stopped. Else if user rate their feelings more than 5 then directly doctors list will be allotted to them by analysing the doctor's data for particular disease.

Module Descriptions:

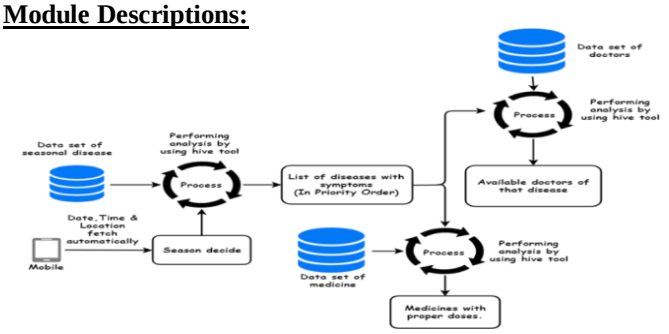


Fig..5: (Automatic)

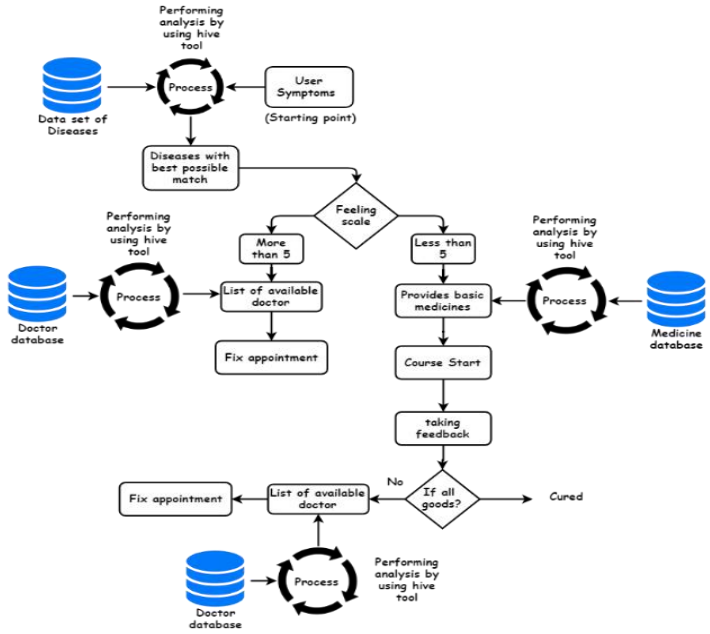
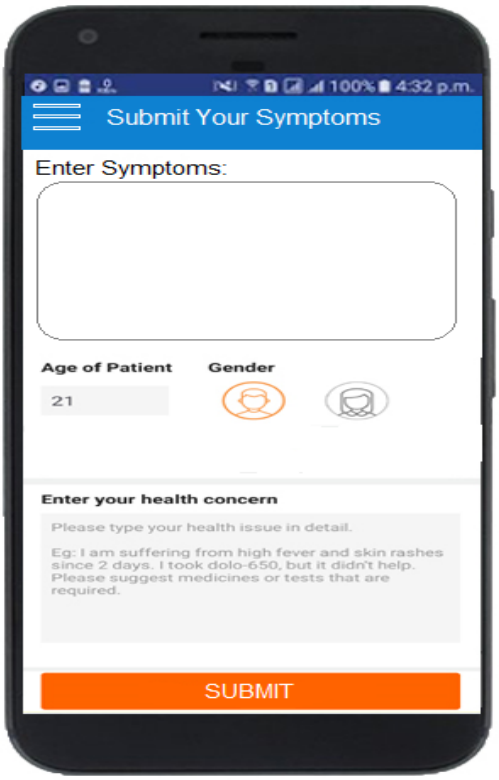


Fig.6: (Manual)

Activity



MANUAL SYSTEM

AUTOMATIC SYSTEM

II. FUTURE SCOPE OF BIG DATA IN HEALTHCARE

- **Improve Hospital Administration:** Big data can help the flow of mundane administrative processes happen in a hassle-free manner. It helps healthcare providers to handle data from diverse sources and provides insights on how to plan for crises like epidemic outbreaks and natural calamities.
- **Fraud Prevention and Detection**: Healthcare environments are also susceptible to human error. The presence Big data helps to prevent a wide range of errors that could happen during storing, sharing and sorting different types of hospital records. Everything from faulty prescriptions and investigations to fraudulent claims of insurance can be curbed in setups that are powered by big data. on the side of health administrators in the form of wrong dosage, wrong medicines, and other human errors. It will also be particularly useful to insurance companies. They can prevent a wide range of fraudulent claims of insurance.

III. CONCLUSION

As we know that disease in city/town/country spread rapidly which result into its poor management and rehabilitation due to unavailability of desired number of doctors and lack of alertness about disease occurrence.

in order to minimize these factors, there is a necessity of an application that can predict about diseases and alert people about them my application exactly focusses on these factors and result into efficient management and rehabilitation of disease as they arrive.

IV. ACKNOWLEDGEMENT

I Would like to thank. Few people who helped me in the preparation of this research Paper. I would like to thank Pankaj Sir of CS Department of Meerut, India.

He encouraged me and provided the knowledge which helped me in the preparation of research paper. I would like to thank Pankaj Sir encouraging me and fulfilling all my need.

V. REFERENCES

[1]. Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. IEEE J Biomed Health Inform 2015;19:1193–1208.
[2]. Archenaa J, Anita EM. A survey of big data analytics in healthcare and government. Procedia Comput Sci 2015;50:408–13.
[3]. Borne K. Top 10 big data challenges – a serious look at 10 big data V's. MAPR, 2014:NO4, 80.
[4]. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PLoS One 2016;11:e0157077.
[5]. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. –Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng2017;64:263–73.
[6]. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Inform Insights 2016;8:1.
[7]. https://www.techopedia.com/30112 /hive
[8]. According to Canalys, In Q2 2009 Android
[9]. https://www.python.org
[10]. Apache Hadoop. Available at http://wiki.apache.org/hadoop.
[11]. Hive wiki at http://www.apache.org/hadoop/hive.
[12]. Hadoop Map-Reduce Tutorial athttp://hadoop.apache.org/common/docs/current/mapred_tutorial.html.
[13]. Hadoop HDFS User Guide athttp://hadoop.apache.org/common/docs/current/hdfs_user_guide.html.
[14]. Apache Thrift. Available at http://incubator.apache.org/thrift.
[15]. Hive Performance Benchmark. Available at http://issues.apache.org/jira/browse/HIVE-396
[16]. Hadoop Pig. Available at http://hadoop.apache.org/pig
[17]. R. Chaiken, et. al. Scope: Easy and Efficient Parallel Processing of Massive Data Sets. In Proc. of VLDB, 2008.
[18]. HadoopDB Project. Available at http://db.cs.yale.edu/hadoopdb/hadoopdb.html
[19]. MicroStrategy. Available at http://www.microstrategy.com
[20]. https://www.djangoproject.com/

Author's Profile

Neha Maurya is pursuing M. Tech. from Subharti Institute of Engineering and Technology, Swami Vivekanand Subharti University, Meerut, India. She received her B. Tech Degree in computer science and Engineering from Uttar Pradesh Technical university, Lucknow, India.

Er. Pankaj Pratap Singh received his B. Tech Degree in Computer Science Engineering from Uttar Pradesh Technical University, Lucknow, India, in 2007 and M. Tech degree in Medical Image and Image Processing from Indian Institute of Technology Kharagpur, Kharagpur, India, in 2010. He is currently working as Assistant Professor in the Department of Information technology, Subharti Institute of Engineering and Technology, Swami Vivekanand Subharti University, Meerut, India. His research interests include IOT, Neural Network, Machine Learning, Deep Learning, Image Processing techniques, Cognitive Science, Computer Network and Data Mining techniques

Er. Amit Kishor is working as Assistants Professor in the department of Computer Science Engineering and I.T., Subharti Institute of Engineering and Technology, Swami Vivekanand Subharti University, Meerut, India. Currently he is pursuing Ph. D. inComputer Engineering from Department of Computer Science and I.T., Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad.

Er. Anirudh kumar Tripathi is working as Assistant Professor in the department of Computer Science Engineering and I.T., Subharti Institute of Engineering and Technology, Swami Vivekanand Subharti University, Meerut, India.