International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Analysis of Multi-Disease & Prediction of Suitable Drug for Healthcare Application using Bigdata

Palash Patidar[1], S. Praneeta[2], Rishabh Kataria[3], B. S Vidhyasagar[4]
[1,2,3,4]*SRM University, Chennai, India*

Abstract: ***The Healthcare technologies have grown immensely in various domains. These technologies have also made the health care data huge making it difficult to process. Several precautions should be taken using pharmaceutical drugs, for both healthcare professionals, who prescribe and administer drugs, and for drug consumers. To achieve this goal, we integrate the massive data using Hadoop and perform a wide range of medical and healthcare functions to infer knowledge that assist in diagnosis of disease and provide the best drug. User can post the query through system. We also arrange appointment to the Best Doctor for the consultation based on user feedbacks.***
Keywords: Big Data; Hadoop; Healthcare; multidiseases analysis; Analytics with Hadoop; Unstructured Medical data,Hive.

I. INTRODUCTION

Traditional drug development is expensive and therefore an alternative method has been invented. Several precautions should be taken while using pharmaceutical drugs, by healthcare professionals, who prescribe as well as consume the drug. The side effects depends on characteristics of patients, such as age, gender, lifestyles, and genetic profiles. All these factors can be a huge problem to professionals and drug consumers

As the enterprises face major issues gathering large chunks of data, they found that the data cannot be processed using any of centralized architecture therefore shifting to distributed architecture. Hadoop is used in analysis of big data as it is Scalable, Cost-effective, Flexible, Fast, and is able to recover from failures easily compared to other methods. This paper describes the scope of big-data in healthcare industry.[1]

Our goal is to provide a tool to help professionals and consumers in finding and choosing the suitable drug. This approach allows a user to query for drugs that satisfy conditions based on drug indications, side effects, and drug interactions,.[2]

A. Role of Big Data Analysis in Healthcare

New discoveries in Big Data analytics will be useful in advanced personalized care, and will help to avoid expenses. The usage of big data, will give different effectiveoutcomes across large population.[3] Surveys of more than a decade can be processed and analyzed to check the diseases that come at a particular period .

B. Role of Hive in Analysis

Hive supports queries expressed in HiveQL, a SQL-like language - which are compiled into map-reduce jobs that are executed using Hadoop. In addition, HiveQL enables users to use custom map-reduce scripts into queries. Hive makes the work a lot easier for data analysis by providing languages that can be easier to implement.[4]

C. Role of Hadoop and Map Reduce

Apache Hadoop has two major components. One is Hadoop Distributed File System (HDFS) and another is Map Reduce. Hadoop/HDFS is for data storage. Map Reduce programming framework uses two tasks common in functional programming: Map and Reduce. The basic idea of Map Reduce is to split the large input data set into many smaller chunks and assign small tasks to multiple devices in a distributed environment. The input files will be automatically split . Later on, the inputs in key-value pair format will be sent to Map function. These input pairs will be processed by map function and generate intermediate key-value pairs which will be inputs for Reduce function. The inputs which have the same key are combined and the final result will be generated by reduce function. The final result will be written into the distributed file system(HDFS) [5].

II. RELATED WORK

There are multiple drug information databases available for public access, such as DrugBank[6] , KEGG DRUG . The author C. Knox et al's database used by professionals and drug consumers in order to make decisions.. There are a few studies that aim to

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

answer medical questions, which include drug-related questions, and provide decision support on drug prescription. There are recent works on predicting drug properties(including drug targets, indications, and adverse effects), which are related to our problem of quantifying the likelihood of missing associations between drugs and drug properties. Methods that predict drug properties based on chemical structures have been proposed.

Samamon Khemmarat, Lixin Gao proposed an approach to query for drugs that satisfy a set of conditions based on drug properties to avoid drug prescription requires consideration of several factors, such as drug interactions and side effects[7].

Charalampos Doulaverakisand George Nikolaidis they proposed a method which provides an efficient drug recommendations service where Semantic Web technologies are coupled with traditional business rule engines[8].

Akhil Langer et al proposed a design of a drug QA system that could be used for providing information about medicines over short message service (SMS). Circulation of medical information using mobile phones is still ina nascent stage because of their limited features – lack of penetration of mobile internet, small screen size etc[9].

JiahuiJin et al's proposed an algorithm for finding the best k answers for a given query without pre-computing graph indices, they further discover found that the answers that match exactly as well as answer that are similar to queries[10].

Asma Ben et al's. approach was to present a complete question analysis including medical entity recognition, semantic relation extraction and automatic translation to SPARQL queries. Designing question answering systems requires efficient and deep analysis of natural language questions[11].

III. ARCHITECTURE AND APPROACH

Most of the related works mentioned before do not depend on the Hadoop techniques that lead to the problem of long retrieval time. Other related works depend on Hadoop technique like mahout that has a slow retrieval time. Our proposed recommendation system architecture delivered overcomes the problem of time delay based on hive and its query language Hive QL as one technique from the Hadoop environment. On the client-server architecture, there is a direct interaction between server and clients. The master server contains the name node; HDFS and slaves that have data nodes that include data and attached storage. On the other hand, on the layered architecture, each layer provides a service to its upper layer and act as a client to the below layer. The first layer contains the hardware devices that are used on the client side. The second layer is the software layer that is divided into nested layers such as HDFS, which is considered the main distributed file system for Hadoop platform. Mapreduce layer built on top of HDFS layer that is used as a framework for easily writing applications that process huge amounts of data in parallel on large clusters. Hive query language is considered the third layer in our software architecture which is used as a data warehouse platform for storing data and it allow users to execute queries using Hive Ql query engine. The recommender layer which is built in the top of the Hive layer contains algorithms applied facilitate to users to search for their need for scientific papers in efficient retrieval time. On the other hand, the interfacelayer is considered the upper layer of this architecture that is used to facilitate the searching process for the users in a friendly user interface.

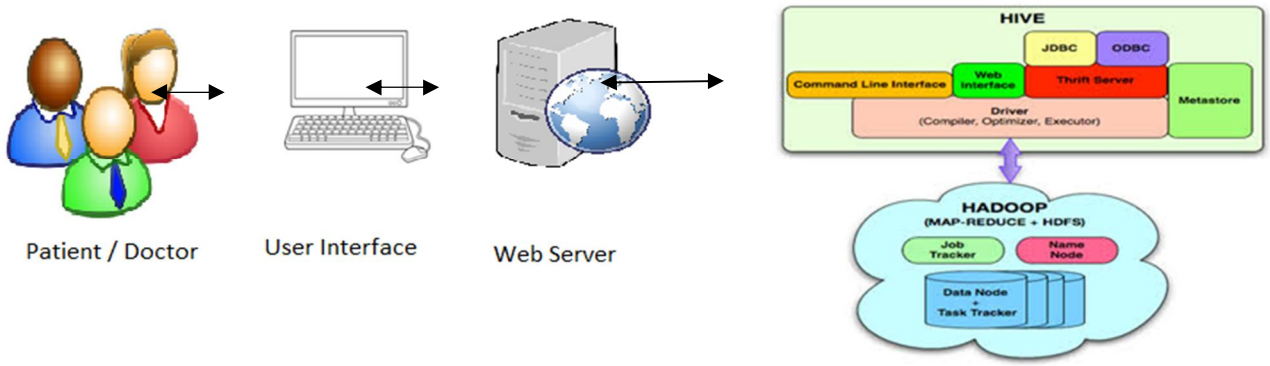

Fig.1 Description of architecture

The above diagram shows the basic Hadoop Hive architecture(hadoop ecosystem)**.** The diagram represents UI in two different forms one is CLI (Command Line Interface),JDBC/ODBC and other is Web GUI (Web Graphical User Interface).This shows when user comes with CLI(Hive Terminal) it directly connected to Hive Drivers, When User comes with JDBC/ODBC(JDBC Program) at that

time by using API(Thrift Server) it connected to Hive driver and when the user comes with Web GUI(Ambari server) [12]it directly connected to Hive Driver. The hive driver receives the tasks(Queries)[13] from user and send to Hadoop architecture.The Hadoop architecture uses name node, data node, job tracker and task tracker for receiving and dividing the work what Hive sends to Hadoop The Patient or doctor can log in using the web user interface[14]. The admin can add multiple disease and their symptoms to train the machine. If the patient logs into system. He can add his symptoms and can request for doctor's appointment[15].

The UI calls the execute interface to the Driver[16].Then Driver creates a session handle which is nothing but request and response mechanism, the query request and sends the query response to the compiler to generate an execution plan. The compiler needs the metadata to send a request for getMetaData and receives the sendMetaData request from MetaStore. This metadata is used to typecheck the expressions in the query tree. The plan generated by the compiler is a DAG of stages with each stage being either a map/reduce job, a metadata operation or an operation on HDFS. For map/reduce stages, the plan contains map operator trees (operator trees that are executed on the mappers) and a reduce operator tree (for operations that need reducers).

The execution engine submits these stages to appropriate components. In each task(mapper/reducer) to read the rows from HDFS files ,the deserializer associated with the table or intermediate outputs is used and these are passed through the associated operator tree. When the output generate it is written to a temporary HDFS file though the serializer. The temporary files are used to provide subsequent map/reduce stages of the plan. For DML operations the final temporary file is moved to the table's location. For queries, the contents of the temporary file are read by the execution engine directly from HDFS as part of the fetch call from the Driver

IV. COMPONENTS OF HIVE

A. *User Interface:* UI means User Interface, it has two forms of UI one is command line interface and other is Web UI. The user interface for users (patient/doctors) to log in and submit queries and other operations to the system.

B. *Driver:* The Driver is used for connecting with UI .Used to get queries using APIs modeled on JDBC/ODBC interfaces. This component implements the notion of session handles and provides execute and fetch

C. *Metastore:* The Metastore is the system catalog which stores all the information in structured form of the various tables in the warehouse and also information about the column type, the core concept of serializer and deserializer is used to read and write data and also the related distributed files where all the data is stored.

D. *Compiler:* The driver calls the compiler with the HQL string which can be one of query statements. One of the component which parses the query, does required semantic analysis on the various query expressions and query blocks which will generates an execution plan using the table and partition metadata from the metastore.

E. *Execution:* Engine For all queries, the data of the temporary files are read by the execution engine directly from HDFS (which is one of the most important component of hive architecture), which will fetch from the Driver. The work of execution engine is to manage the dependencies between the different stages of the plan and execution of these stages takes place on the system components.

V. CONCLUSION

In this paper, we propose an approach for answering drug queries to support drug prescription. Our focus is on how to obtain and rank answers based on incomplete information and provide personalization. To cope with incomplete and noisy data, we allow both exact and close matches when answering queries. We will be using Hadoop Framework and Hive queries to process large dataset which will provide relevant results.

REFERENCES

[1] JyotiNandimath, Ekata Banerjee, AnkurPatil "Big data analysis using Apache Hadoop" 2013 IEEE
[2] SamamonKhemmarat and LixinGao "Predictive and Personalized Drug Query System" 2016 IEEE
[3] Asif Adil,Hushmat Amin Kar "Analysis of Multi-diseases using Big Data for improvement in Healthcare" 2015 IEEE
[4] Hongyong Yu, Deshuai Wang "Data Management and Analysis Based on Hadoop and Hive"2012 Conference on Computational and Information Sciences
[5] Jun Ni, Ying Chen, JieSha, and Minghuan Zhang"Hadoop-based Distributed Computing Algorithms for Healthcare and Clinic Data Processing" 2015 International Conference on Internet Computing for Science and Engineering.

[6] C. Knox et al., “Drugbank 3.0: a comprehensive resource for omicsresearch on drugs,” Nucleic acids research, vol. 39, no. suppl 1, pp.D1035–D1041, 2011.
[7] SamamonKhemmarat, LixinGao “Supporting Drug Prescription via Predictive and Personalized Query System” Journal of biomedical and health informatics, december 2015.
[8] CharalamposDoulaverakis, George Nikolaidis, Athanasios Kleontas MD, IoannisKompatsiaris “A Semantic-enabled Framework for Drug Recommendations” Journal of Biomedical Semantics,2014.
[9] AkhilLanger,RohitBanga,AnkushMittal,L.V. Subramaniam,ParikshitSondhi“A text based drug query system for mobile phones”Int. J. Mobile Communications
[10] JiahuiJin, SamamonKhemmarat, Lixin Gao, Junzhou Luo “Querying Web-Scale Information Networks Through Bounding Matching Scores”IW3C2, 2015.
[11] Asma Ben Abacha,PierreZweigenbaum “Medical Question Answering:Translating Medical Questions into SPARQL Queries” IHI’12, January 28–30, 2012.
[12] A. Langer et al., “A text based drug query system for mobile phones,”Int. J. Mob. Commun., vol. 12, no. 4, pp. 411–429, Jul. 2014.
[13] A. Ben Abacha and P. Zweigenbaum, “Medical question answering:translating medical questions into sparql queries,” in Proceedings of the 2nd ACM SIGHIT. ACM, 2012, pp. 41–50.
[14] M. Kuhn et al., “A side effect resource to capture phenotypic effects ofdrugs,” Molecular systems biology, vol. 6, no. 1, p. 343, 2010.
[15] A. Pavloet. Al. A Comparison of Approaches to Large-Scale Data Analysis. Proc. ACM SIGMOD, 2009.
[16] M. Dumontier and N. Villanueva-Rosales, “Towards pharmacogenomicsknowledge discovery with the semantic web,” Briefings in bioinformatics,vol. 10, no. 2, pp. 153–163, 2009.