Analysis of Big Data in Healthcare and Life Sciences Using Hive and Spark

A. Sai Hanuman, R. Soujanya and P. M. Madhuri

Abstract Big data is a declaration used to recognize the database whose area is afar the potential of typical database software tools to store, organize and examine. Big data has shown a new path toward the mankind. With several theoretical and technological obstacles in health huge processing, it is onerous to transfer knowledge into fortunate and valuable applications. Meeting the challenge of handling big data in healthcare information construction procedure, this paper proposes a referential architecture on the Hive and Spark platform to overcome the problems in healthcare big data process. Hive is a noteworthy project as a result of it permits exposing the simplest components of Hadoop, specifically map reduce and knowledge storage. Spark may be a memory-based computing framework that features a higher ability of computing and fault tolerance, supports batch, interactive, iterative and flow calculations. Experiment results of data upload, data query and data analysis show that the performance of the proposed framework is greatly improved, and a brief summary of the performance and the differences between two methods of Hive and Spark is also discussed.

Keywords Big data · Hadoop · Healthcare · Hive · Spark and Scala

A. Sai Hanuman · R. Soujanya · P. M. Madhuri (✉)
Department of Computer Science and Engineering, GRIET, Bachupally, Hyderabad, Telangana, India
e-mail: pmmadhuri18@gmail.com

A. Sai Hanuman
e-mail: a_saihanuman@hotmail.com

R. Soujanya
e-mail: soujanya96@gmail.com

K. S. Raju et al. (eds.), *Data Engineering and Communication Technology*,
Advances in Intelligent Systems and Computing 1079,
https://doi.org/10.1007/978-981-15-1097-7_69

1 Introduction

1.1 Big Data and Hadoop

Big data: Big data is "a data that is in huge size, high speed and variable data that wants innovative techniques and new technologies to enable the capture, distribution, management, memory and analysis of the information."

Some characteristics of big data are:

- **Volume**: It is the amount of data produced by organizations or individuals. All organizations are searching for ways to handle the ever-increasing data volume that is being created every day [2].
- **Velocity**: It is the frequency and speed at which data is captured, produced and shared. Consumers as well as businesses now producing lots of data and in shorter cycles, from hours, minutes, seconds down to milliseconds.
- **Variety**: It is the creation of new data types together with social, mobile and machine resources. New types include metrics, content, physical data points, mobile, process, location or geo-spatial, machine data, radio-frequency identification (RFID), hardware data points, search and web. It also includes unstructured data.
- **Veracity**: It is outlined because of the exactness of information. Incorrect knowledge will cause tons of issues for organizations. Hence, organizations want to ensure that the data is correct and analyses performed on the data are precise. In robotic decision-making, where no human is involved we need to be sure that both the data and the analyses are correct [7, 12] (Fig. 1).

Hadoop 2.0. An open-source software framework is provided by Apache software foundation, which is used store and processing the large-scale data sets with clusters of commodity hardware. It is designed for scaling up from a single server to

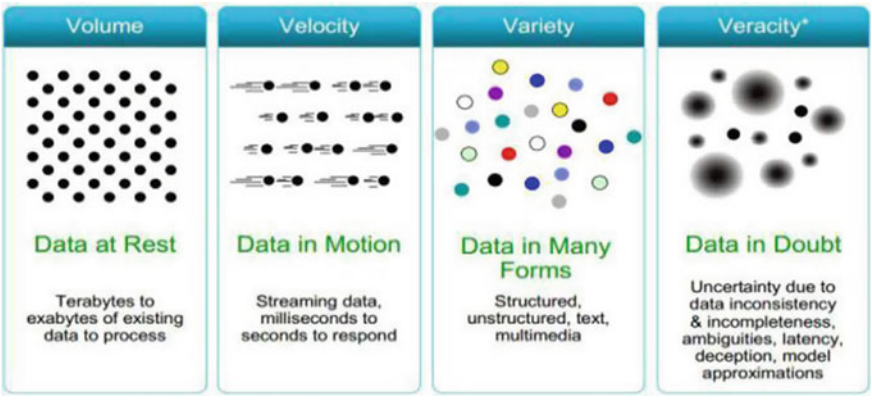

Fig. 1 Big data

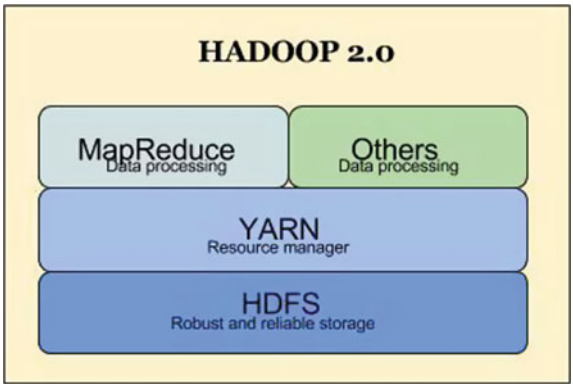

Fig. 2 Hadoop framework

thousands of servers in a cluster with local computation and storage on every individual server. There are three main core components inside this run-time environment (from bottom to top) [1] (Fig. 2).

1.2 Big Data Analytics

Big data analytics is a strategy, for analyzing giant volumes of information. This huge knowledge is gathered from a good style of sources, including sensors, social networks, videos, sales transaction records and digital images. The aim of analyzing all this data is to find out new patterns and connections that may somewhat be imperceptible and that might offer valuable insights about the users who created it.

1.3 Types of Analytics in Big Data Analytics

There are four types of analytics. Here, we tend to begin with the only one and go all the way down to a lot of subtle. As it happens, the more complex an analysis is, the more value it brings (Fig. 3).

- **Descriptive** analytics, which uses data aggregation and data mining to provide insight into the past and answer: "What has happened?"
- **Diagnostic** analytics, which uses historical data, can be measured against other data to answer the question of "why something happened?"
- **Predictive** analytics, which uses statistical models and forecasts techniques to understand the future and answer: "What could happen?"
- **Prescriptive** analytics, which uses optimization and simulation algorithms to advice on possible outcomes and answer: "What should we do?" [13]

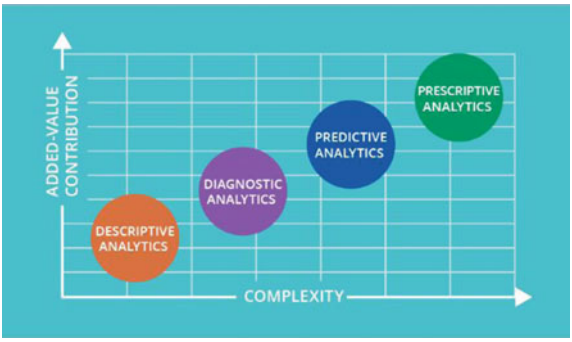

Fig. 3 Types of analytics

2 Big Data Analytics in Healthcare

Regular methods can be changed in any industry by using big data. By applying data analytics in healthcare industry, we can reduce the cost for treatments, can prevent avoidable diseases and quality of life can be improved. Based on treatment methods using nowadays, challenges are also increasing as there is continuous increment in the population count. Just like business people, healthcare professionals are also accumulating the information of patients and are using those data for the research.

By gathering large volume of information in healthcare, how it is going to help? We can show huge knowledge examples in attention that exist already which we tend to take pleasure in [14].

2.1 How to Analyze Large Data in Healthcare

- There are many positive and lifesaving outcomes when we apply big data analytics in healthcare each and every information creates huge volume of data that is used by specific technologies to get analyzed. Application toward healthcare data gets facilitated by avoiding cost issues and early detection.
- Now that we tend to live more, treatment models have altered and a lot of those progressions are explicitly determined by data. Specialists need to get a handle on the greatest sum as they will a couple of patients and as right off the bat in their life as potential, to pick up notice indications of critical sick—treating any ailment at a beginning period is way a ton of clear and less pricy. With medicinal service data investigation, the obstacle is best than fix and figuring out how to draw a far-reaching picture of a patient can give protections, a chance to offer a custom-fitted bundle. This is regularly the business's imagine to handle the storehouse's issues a patient's data has: Everyplace are gathered bits and

bytes of it and filed in emergency clinics, centers, medical procedures and so on with the difficulty to talk appropriately.

- A considerable length of time gathering tremendous measures of information for therapeutic use has been expensive and long. With the present continually enhancing advances, it winds up less demanding not exclusively to accumulate such data anyway conjointly to change over it into important pivotal bits of knowledge that may then be acclimated offer higher consideration. This is regularly the point of medicinal service data investigation: abuse information-driven discoveries to foresee and understand a pull before it is past the point of no return, anyway conjointly evaluate procedures and covers speedier, monitor stock, include patients a great deal of in their very own well-being and engage them with the instruments to attempt [14].

2.2 *Analyzing the Data Flow in Healthcare*

The business is growing fast and it has several categories of information which are setting out to remodel it—in the other hand, there is still availability of labor and the field slowly implements the different technologies that may be helpful in the long run and for effective business operations. The below are some of the examples laid out to analyzing the data in healthcare.

- Analyzing the patients' footprint will helpful to recruitment staff accordingly.
- Usage of Electronic Health Records (EHRs).
- Periodic alerts for health checkup.
- Improvise research and development to cure cancer.
- Usage of predictive analytics.
- Help in averting narcotic maltreatment in the USA.
- Improve patient awareness in their own health.
- Use health information for a better-informed strategic coming up with.
- Improve the data security.
- Exercise telemedicine.
- Fit in medical imaging for broader diagnosis.
- Avoid needless ER visits.

These samples of big data in medicinal services demonstrate that the event of medical applications of information ought to be the apple inside the eye of knowledge science as they require the possibility to spare heaps of money and most essentially, individuals' lives. As of now these days, it permits for early identification of sicknesses of individual patients and financial groups and accepting preventive activities as a result, we have a tendency to all understand, hindrance is best than cure [14].

2.3 *The Most Effective Method to Use Big Data in Healthcare*

All things considered, these applications of big data in healthcare 3 principle patterns: the patient's ability may enhance drastically, together with best quality of medication and satisfaction; the overall healthcare of the residents should improved after sometime; and thus, the general cost should be reduced. We have a look for best approach to utilize enormous information in medicinal services, in a hospital for case [14] (Fig. 4):

This healthcare dashboard furnishes you with the outline required as a hospital director or as a facility manager. Assembling in one focal reason all the data on each division of the hospital, the gathering activity, its temperament, the costs acquired, etc., which is a good facilitate to run it swimmingly.

Here, the foremost necessary principal regarding varied aspects: The quantity of patients existed in your facility, on the other hand they stayed long time, how much cost to check patients, and the waiting time in crisis rooms. Such a holistic read helps top administration to determine potential bottleneck and patterns over time. This key will improve the common operations performance, improving the patient treatment and having the exact requirements for good staffing.

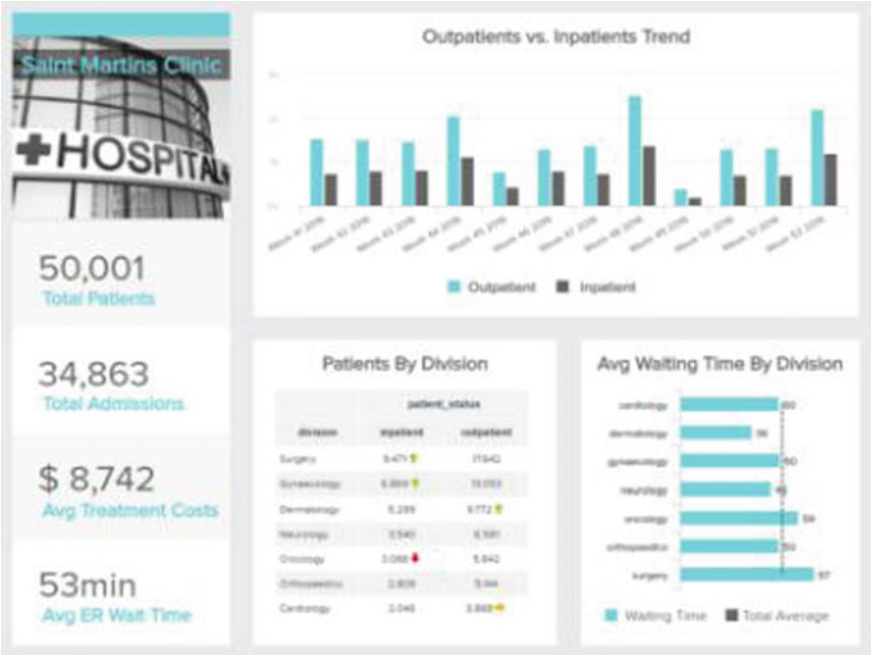

Fig. 4 Usage of data analytics in healthcare in a hospital

3 Diverse Tools to Analyze Big Data in Healthcare

Hadoop Ecosystem is an open-source framework by Apache software foundation written in Java for processing and storing large volumes of data with clusters of commodity nodes which solve big data problems. To analyze the huge data in less time and choose the right tool, here some of top big data tools in the areas of data mining, storing, integrating, cleaning, visualizing, extracting and analyzing [15].

- **HDFS (Hadoop Distributed File System)**: This is a core component of Hadoop framework and used for distributing data across a cluster of commodity hardwares.
- **Apache Sqoop**: A tool designed to transferring data between Hadoop and other relational database servers.
- **Apache Flume**: It is a framework which will efficiently collecting, aggregating, moving large amount of streaming data into HDFS.
- **Apache HBase**: A column-oriented database model which runs on top of Hadoop.
- **Apache Pig**: A platform/tool for analyzing massive amount of data in Hadoop in parallel.
- **Zookeeper**: A admin tool for managing the jobs in Hadoop clusters.
- **Apache Hive**: It is an open-source data warehouse, and it uses language called HiveQL which is similar to SQL.
- **Apache Ambari**: A tool which is responsible for keeping track of applications and their status.
- **Apache Spark**: A new way of running algorithms even faster than Hadoop.
- **Apache Mahout**: It is open-source framework, primarily used for creating machine learning algorithms.

4 A Conceptual Architecture of Big Data Analytics

The abstract structure for a big data analytics project is a standard health analytics project. In this health analytics project, a business intelligence platform will perform the analysis of data and put data into a complete system, like a portable computer, or a desktop. By definition, big data is data it is large in volume, variety and velocity and executes this massive data across multiple nodes. The distributed computing process has occurred for decades. Use terribly large data sets for analyzing new patterns and achieve big insight for creating better health-related decisions. Moreover, some of the open-source platforms like Hadoop or map reduce, available on the cloud, for developing any new application for analyzing massive data in healthcare [3, 4] (Fig. 5).

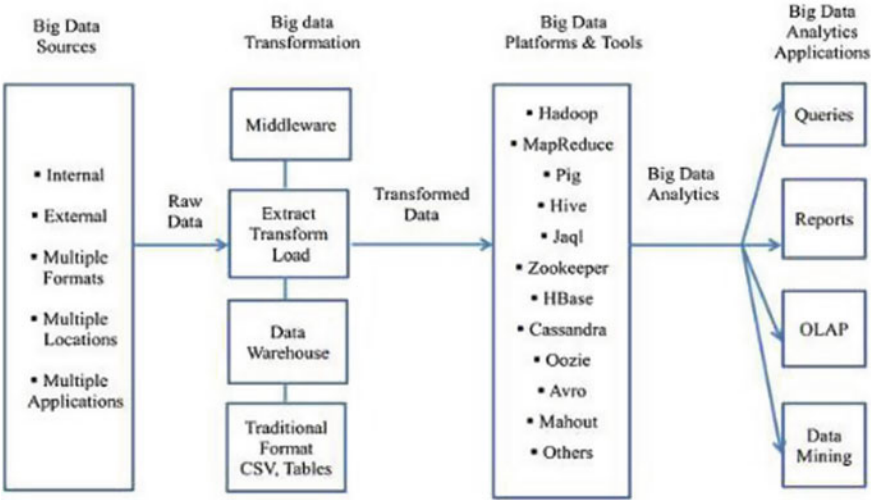

Fig. 5 Conceptual architecture of big data analytics

5 Methodology and Implementation

This paper described healthcare data set at first, later about the analysis of data provided using various tools like Hive and Spark. Then, discussed how efficiently the data will be processed.

Hive is an open-source data warehouse system designed on top of Hadoop. It uses language called HiveQL (HQL) which is similar to SQL. HQL is used for querying, analyzing table data which are put in storage as flat files on Hadoop Distributed File System. HQL automatically translates SQL like queries into a group of map reduce jobs which will execute on a Hadoop Cluster. Hive supports data partitioning and schemas to keeps the metadata in a relational database. Basically, Hive supports partitioning the table on a precise dimension. For instance, we could store patient particulars and partition the table on disease info. This allows for creating queries in an organized data model in the future [11].

Spark is a memory-based computing framework model, and it supports fault tolerance iterative, batch processing, interactive, and flow calculations. It absorbs the benefits of Hadoop Map Reduce model, but unlike Map Reduce, the data is kept in random access memory (RAM) instead of some slow disk drives and is processed in parallel, it is called as Memory Computing. This will improve the efficiency of data computing [8, 9].

5.1 Implementation

Healthcare data sets:

The data consists of different patient details which are used to analyze data in healthcare [6, 16]. The data set attributes are:

1. **Patient_Pid**: Unique identification of patients
2. **Patient_Pname**: Name of the patient
3. **Age**: Patient's age at diagnosis
4. **Gender**: Patient gender
5. **Disease_Info**: Type of the disease to analyze
6. **Hospital_Name**: Name of the hospital for treatment
7. **Admited_Date**: Patient admitted in hospital
8. **Address:** For communication (Fig. 6)

Below are some implementation steps for analyzing healthcare data set [10].

Step 1: Collect different disease information of the patients from various hospitals.

Step 2: Convert the data set into .csv format and copy the file to cloud X lab.

Step 3: Create a table using Structured Query Language, i.e., SQL.

```
Query: create table patient_details (Patient_Idstring,
Patient_Pname string,AGE int, GENDERvarchar (6),
DISEASE_INFO string, HOSPITAL_NAME string, ADMITED_DATE
varchar(15), ADDRESS string);
```

Step 4: Load the data into MYSQL.

	PATIENT_ID	PATIENT_NAME	AGE	GENDER	DISEASE_INFO	HOSPITAL_NAME	ADMITED_DATE	ADDRESS
2	100001	Manish_Jain	54	Male	Typhoid-Fever	FRESENIUS-MEDICAL-CARE	10/21/2015	MOHULIA-BARHAMPURA, P.S.-BARHAMPURA, MUZAFFARPUR, Bihar, INDIA, 842001
3	100002	A.Venketeshwara_Reddy	65	Male	Cancer	GAMBRO-HEALTHCARE	12/11/2015	TILAK_NAGAR_WARD_NO.10_BEGUSARAI_Bihar_INDIA_851101
4	100003	G.Banu	76	Male	Cancer	GAMBRO-HEALTHCARE	1/9/2015	45,Raja_Bazar_Anchal_-_Motihari_Motihari_Bihar_INDIA_845401
5	100004	P.Srinivasa_Rao	67	Female	Typhoid-Fever	FORTIS_HOSPITAL	1/4/2015	WARD_NO.-23,A.D.NAGAR,BADHARGHAT,AGARTALA_Tripura_INDIA_799001
6	100005	Adil_Khan	67	Female	Cancer	GAMBRO-HEALTHCARE	2/23/2015	PADUMONI_A.T.ROAD_(NH-37),NEAR_KUTUHABORIA_TINALI_JORHAT_Assam_INDIA_7850
7	100006	Sri_Bad	56	Female	Typhoid-Fever	FRESENIUS-MEDICAL-CARE	3/3/2016	OLD_NO.4,NEW_NO.6,FIRST_FLOOR_100_FEET_ROAD,ELLAIPILLAICHAVADY_PUDUCHERRY_6
8	100007	Gandhe_Krishna	45	Female	Cancer	GAMBRO-HEALTHCARE	6/4/2015	C/O_BABLU_KUMAR,MAIN_ROAD_BIRPUR_WARD_NO-8,PO_AND_PS_-BIRPUR,DISTT_-SUPA
9	100008	P_V_Raju	34	Female	Cancer	FRESENIUS-MEDICAL-CARE	9/27/2015	AWADESH_PD_SINGH_C/o_Narayan_Singh,Awadhpuri,Digha_PATNA_Bihar_INDIA_800011
10	100009	Shakti_Sagar	18	Female	Genetic_Disorder	FRESENIUS-MEDICAL-CARE	10/6/2015	AKRIYAPUR,TANSPORT_NAGAR,APPOSITE_GATE_NO.-2_PAHARI_PATNA_Bihar_INDIA_80
11	100010	V.A.Rao_Aevas	26	Female	Liver_Disorder	FRESENIUS-MEDICAL-CARE	11/8/2015	Manpur,Kumhar_Toli,P.O.-Buniyadganj_Moffasil_Gaya_Bihar_INDIA_823001
12	100011	A.V.S.N.Sarma	17	Female	Typhoid-Fever	FRESENIUS-MEDICAL-CARE	1/29/2016	Flat_No.204,Krishna_Apartment,Puran_Vihar_Argora_Bypass_Road_Argora,P.O.-Ashok_Nag
13	100012	S.Prabhakar_Rao	46	Female	Cancer	NATIONAL-RENAL-INSTITUTES	8/23/2016	LATE_PADAMSHREE_BHARAT_MISHRA_LANE_NEW_COLONY_PAKRI_ARA_Bihar_INDIA_8023
14	100013	P.G.P.Vittal	56	Male	Heart_Disorder	GAMBRO-HEALTHCARE	1/27/2015	TIRUPATI_TOWER,CIRCULAR_ROAD_PO_&_PS:_LALPUR_RANCHI_Jharkhand_INDIA_834001
15	100014	Joy_Paulo	75	Male	Cancer	FRESENIUS-MEDICAL-CARE	10/23/2015	HOUSE_NO.57_BASISTHAPUR,BYE_LANE_4,SURVEY,BELTOLA_GUWAHATI_Assam_INDIA_781
16	100015	M.J.Betala	84	Female	Genetic_Disorder	FRESENIUS-MEDICAL-CARE	1/5/2016	289,NIHAL_HOUSE,RAHAT_COLONY_KAJAMANI_ROAD,NEAR_KABIR_CHOWK_KISHANGANJ
17	100016	B.Sarat_Chand	85	Female	Cancer	DIALYSIS-CLINIC,-INC.	1/5/2016	SHANTI_PATEL_BHAWAN_WARD_NO.-7,MURLIYA_CHAK_SITAMARHI_Bihar_INDIA_843302
18	100017	Srinivas_	28	Female	Cancer	FRESENIUS-MEDICAL-CARE	1/5/2016	C/O-HARI_MADHAN_PRASAD,S/O-SATYA_NARAYAN_PRASAD_CHANIYAPURI,RAJA_BAZAR
19	100018	K.V.Satyanarayana	27	Male	Liver_Disorder	FORTIS_HOSPITAL	1/5/2016	Kashyap_Complex,Ashok_Nagar_Road_No.4,Distt_-Ranchi_Ranchi_Jharkhand_INDIA_834001
20	100019	M.Ameenj	26	Male	Typhoid-Fever	FRESENIUS-MEDICAL-CARE	1/5/2016	H.No.1LL591/Mohalla_-Purandar_Bigha,Sichai_Colony,P.O.-Jupla_PS_-Hussainabad,Palamu
21	100020	K_Padmanabham	36	Female	Typhoid-Fever	FRESENIUS-MEDICAL-CARE	8/29/2016	TDSH_APARTMENT,D-SECTOR_NAHARJAGUN_Arunachal_Pradesh_INDIA_791110
22	100021	N.Khara	47	Female	Cancer	RENAL-CARE-GROUP-INC.	1/5/2016	G/B_HILL_CROWN_APARTMENTS,GR.FLOOR,COLLEGE_ROAD,BARDEZ_MAPUSA_Goa_INDIA_
23	100022	B.K.Neotia	56	Female	Heart_Disorder	DIALYSIS-CLINIC,-INC.	7/23/2015	MTS_199_BISTUPUR_MARKET_JAMSHEDPUR_Jharkhand_INDIA_831001
24	100023	Guda_Krishnaprasad	49	Male	Cancer	RENAL-CARE-GROUP-INC.	1/1/2015	HOUSE_NO.57_BASISTHAPUR,BYE_LANE_4,SURVEY,BELTOLA_GUWAHATI_Assam_INDIA_781

Fig. 6 Different patient details

Query: Load local file 'project/data/dataset.csv' into table patient_details FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' ignore 1 lines;

Step 5: Import data from MYSQL to HDFS using SQOOP.

Query: sqoop import –connect jdbc:mysql://ip-172-31-20-247/sqoopex –username sqoopuser –password NHkkP876rp –table patient_details –target-dir project/healthcare/data/input/stage -m 1 –direct

Step 6: Execute commands in HUE Environment.

6 Results and Analysis

6.1 Using Hive

As we discussed earlier, Hive environment is used for the purpose of analyzing the data. In Hive, the patients' data set should be first loaded into it. The uploaded file is simply a comma separated file. Below figures show how the raw data is uploaded into Hive.

Step 1: Create table in Hive and load the patient details from HDFS using Query:

```
create table patient_details(Patient_Id string, Patient_Pname string, AGE int,GENDER varchar(6), DISEASE_INFOstring, HOSPITAL_NAME string, ADMITEDDATE varchar(15), ADDRESS string);
```

Once the file is uploaded into a Hive environment, analysis can perform on the given data set. The given data set has 2999 records and eight attributes (Fig. 7).

Step2: Analysis of data in Hive without partitioning

Without partition in Hive, the query will give the result within 6.60 s for analyzing the cancer disease info. Analyze the patient details by using Query:

```
select HOSPITAL_NAME, count(DISEASE_INFO) as cancer_count from patient_details where DISEASE_INFO = 'Cancer' group by HOSPITAL_NAME order by cancer_count desc;
```

(Fig. 8)

Step 3: Create one more table in Hive with partition method using below Query:

```
create table patient_data (Patient_Id string,Patient_Pname
string, AGE int, GENDER varchar(6), HOSPITAL_NAME
string, ADMITED_DATE varchar(15), ADDRESS string) partitioned by (DISEASE_INFO string) location 'hdfs://ip-172-31-35-
```

141.ec2.internal:8020/user/mbabu60499425/project/healthcare/data/output/disease';

Step 4: Load patient details into patient_data table using Query:

```
set hive.exec.dynamic.partition.mode = nonstrict;
insert into table patient_data partition (disease_info)
```

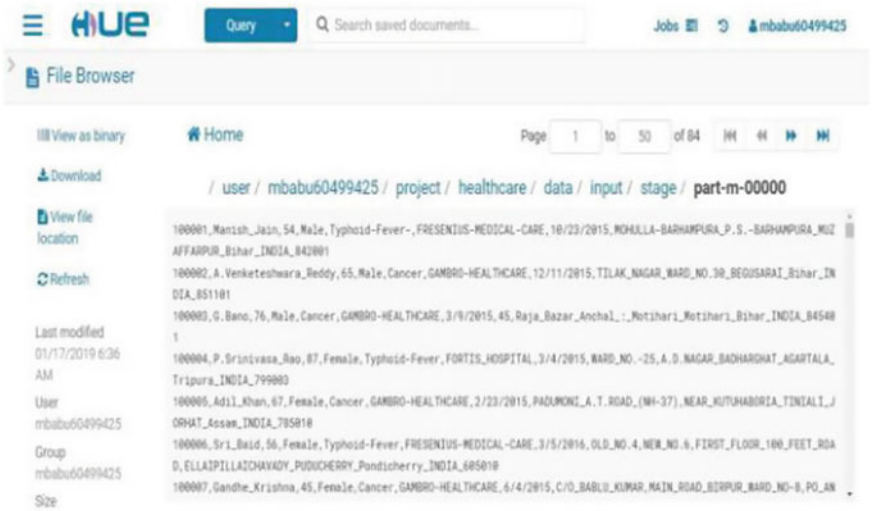

Fig. 7 Loading of raw data to Hive

Fig. 8 Query without partition in Hive: number of cancer patients visited to hospital

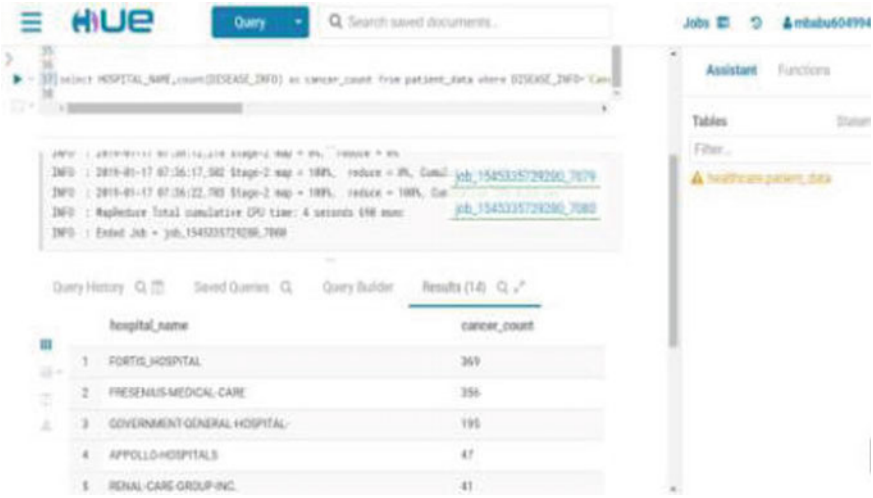

Fig. 9 Query with partition in Hive: number of cancer patients visited to hospital

```
select patient_id, patient_name, age, gender,
hospital_name, admited_date, address, disease_info from
patient_details;
```

Step 5: Analyze the partitioned table using below Query:

```
select HOSPITAL_NAME,count(DISEASE_INFO) as cancer_-
count from patient_data where DISEASE_INFO = 'Cancer' group
by HOSPITAL_NAME order by cancer_count desc;
```
(Fig. 9)

The time taken in Hive environment to perform this analysis was recorded as 4.690 s.

6.2 Using SPARK

In Spark, the results are usually held on in memory rising potency of data computing. It also supports seamless data sharing between applications hence best suited for batch processing, ad hoc querying and streaming processing applications.

6.2.1 Spark SQL

Spark SQL may be a portion on top of Spark Core that presents a new data abstraction denoted as Schema RDD, which is responsible for structured and semi-structured data.

6.2.2 Spark Streaming

Spark Streaming influences Spark Core's for fast planning to perform streaming analytics. It ingests data in small batches and completes RDD (Resilient Distributed Datasets) transformations on those small batches of data.

6.2.3 Machine Learning Library

MLlib may be a distributed machine learning framework higher than Spark. Spark MLlib is nine times faster than Hadoop disk-based version of Apache mahout.

6.2.4 GraphX

GraphX may be a distributed graph processing framework on top of Spark. It provides an API for expressing graph computation that can model the user defined graphs by using Pregel abstraction Python.

6.2.5 Scala

Many of the high-performance data science frameworks will build on top of Hadoop usually which are written and use Scala or Java. Scala has amazing concurrency support. It also runs on the JVM, which makes it almost a no-brainer when paired with Hadoop. Similar to Java, Scala is an object oriented, and uses curly brace syntax like C programming language. Scala has many features of functional programming languages like Scheme, including currying, Standard ML and Haskell, immutability, lazy evaluation, type inference and pattern matching [5].

Here, we discussed some of the steps to analyze the data in spark.

Step1: Import Hive context to Spark and Load data into patient_input using Query:

```
import org.apache.spark.sql.hive.HiveContext
val hiveContext = new org.apache.spark.sql.hive.Hive
Context(sc) import hiveContext._
valpatient_input = sc.textFile
("/user/mbabu60499425/project/healthcare/data/input/
stage
")
```

Step 2: Analyze the data using Query:

```
val cancer_hospitals = hiveContext.sql("select hospi-
tal_name,count(patient_id) as patient_count from health-
care.patient_details where disease_info = 'Cancer' group by
hospital_name order by patient_count desc"); spark.time(-
cancer_hospitals.show)
```

(Fig. 10).

Fig. 10 Query in Spark: number of cancer patients visited to hospital

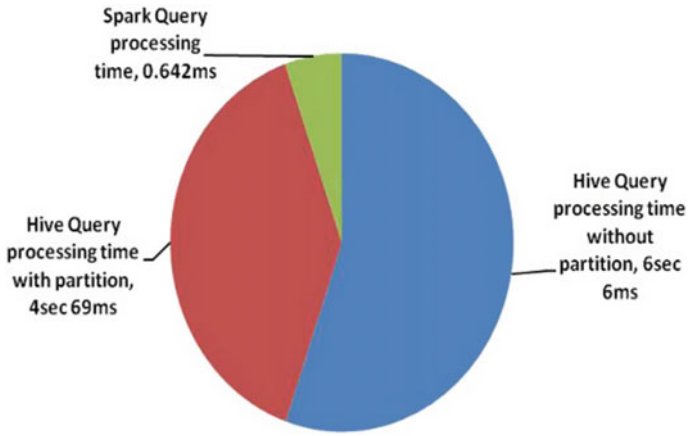

Fig. 11 Pie chart for Query processing time using Hive and Spark

Table 1 Time taken to process the healthcare data in Hive and Spark

S. No	Time (in s)	
Hive	6.06 (without partition)	4.69 (with partition)
Spark	0.642 (in memory)	

Spark will give the result within 0.642 s for analyzing the cancer disease info. **Results**: In this exploration, it is observed that for processing the different patients' data with Spark tool taking less time than Hive (Fig. 11 and Table 1).

7 Conclusion

Big data analytics in healthcare will become a capable field for providing insight from very big data sets and improving outcomes on reducing costs. Its potential is so great; however, there also remain challenges to overcome.

Big data is a collection of data elements whose size, speed, complexity required to adopt software and hardware mechanisms to successfully store, analyze and visualize data. Healthcare is an important example to show how four Vs of data velocity, veracity, variety and volume are vital aspects of the data it produces. The data spread among multiple healthcare centers should provide a platform for global data transparency. Researchers are reviewing the complexity in healthcare data in terms of both characteristics of the data itself and the taxonomy of analytics that can be expressively performed on them. The goal of using Hive and Spark in healthcare is to collect and analyze data from public health trends in a region of people to identify better hospital options for each patient.

References

1. Zaveri, C.: Use of big-data in healthcare and lifescience using Hadoop technologies. Copyright. 978-1-5090-3239-6/17/$31.00©2017IEEE
2. Bhosale, H.S., Gadekar, D.P.: A Review paper on big data and hadoop. Int. J. Sci. Res. Publ. **4**(10),1 (2014). ISSN 2250-3153
3. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Info. Sci. Syst. **2**, 3 (2014)
4. Parimala, S.: A survey on security and privacy issues of big data in healthcare industry and implication of predictive analytics. Int. J. Inn. Res. Comput. Commun. Eng. 0504098. https://doi.org/10.15680/ijircce.2017
5. JayaLakshmi, G., Srisaila, A., MadhaviLatha, P.: Enhancment of healthcare outcomes using big data analytics. IJLTET **7**(3) (2016). ISSN:2278-621X
6. Panda, R.P., Barik, P.P., Prusty, P.A.K.: A review paper on big data in lung cancer big data analytics in lung cancer. Int. J. Trend Res. Develop. **3**(5) (2016). ISSN: 2394-9333 IJTRD
7. Nandhini, S.G., Nandhini, V., Lavanya, K., Kokilam, V.: Big data analytics in health care. IJIRT 2(4), (2015). ISSN: 2349-6002
8. Shrutika Dhoka, R., Kudale, A.: Use of big data in healthcare with spark. Int. J. Sci. Res. (IJSR). ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14| Impact Factor (2015): 6.391
9. Liu, W., Li, Q., Li, X.: A prototype of healthcare big data processing system based on spark. In: 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI). https://doi.org/10.1109/bmei.2015.7401559
10. Durga Sri, B., Nirosha, K., Padmaja, M.: Healthcare Analysis Using Hadoop. **4**(6), 2017. ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697
11. Sadhana, S., Shetty, S.: Analysis of diabetic data set using Hive and R. Int. J. Emer. Technol. Adv. Eng. **4**(7) (2014). ISSN 2250-2459, ISO 9001:2008 Certified Journal. Website: www.ijetae.com

12. https://www.scribd.com/document/107279699/Big-Data-in-Healthcare-Hype-and-Hope
13. https://www.scnsoft.com/blog/4-types-of-data-analytics
14. https://www.datapine.com/blog/big-data-examples-in-healthcare/
15. https://www.kdnuggets.com/2014/08/18-essential-hadoop-tools.html
16. https://www.kaggle.com/rajr16/diseaseinfo