International Journal of Engineering & Technology, 7 (3.29) (2018) 88-90

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Analyzing Indian healthcare data with big data

R. Venkateswara Reddy [1] *, Dr. D. Murali [2]

[1] *Assistant Professor, Department of CSE, CMR College of Engineering & Technology, T.S*
[2] *Professor & HOD, Department of CSE, Vemu Institute of Technology, Tirupathi*
**Corresponding author E-mail: venkatreddyvari@gmail.com*

Abstract

Big Data is the enormous amounts of data, being generated at present times. Organizations are using this Big Data to analyze and predict the future to make profits and gain competitive edge in the market. Big Data analytics has been adopted into almost every field, retail, banking, governance and healthcare. Big Data can be used for analyzing healthcare data for better planning and better decision making which lead to improved healthcare standards. In this paper, Indian health data from 1950 to 2015 are analyzed using various queries. This healthcare generates the considerable amount of heterogeneous data. But without the right methods for data analysis, these data have become useless. The Big Data analysis with Hadoop plays an active role in performing significant real-time analyzes of the enormous amount of data and able to predict emergency situations before this happens.

Keywords: *Big Data, Healthcare; Hadoop; Pig Latin; HDFS; Map Reduce.*

1. Introduction

We live in an extremely dependent and data-driven world. There has been a wave of data over the years. 90 percent of global data has been generated in the last two years; more data has been generated in the past two years than in the entire history of mankind. The data grow considerably, by 2020 it is expected that each person will produce an average of 1.7 MB of data per second at the same time. Aggregated collected data 4.4 Zettabytes currently generates 44 Zettabytes, or 44 billion Gigabytes by 2020. At that time, we will have more than 6.1 billion smartphone users, 50 billion connected devices on the Internet and total data collected, of which three in the cloud being saved. At the moment all collected data, only 0.5% is really studied and used, here the enormous potential of data analysis. The advent of social media and smartphones led to increased collection of data. According to an IBM survey around 70 percent of the companies are not prepared to handle the Big Data challenges happening in the market (IBM Survey, 2017) [1].

Big Data is huge volumes of data which cannot be stored and processed using theconventional systems. Due to the large volume it is arduous to achieve effectual analysis using the long established approaches. Big Data posed challenges with its four Vs, properties volume, velocity, veracity and variety. (Katal, Avita, et al, 2015) [2].

Volume: With advent of social media and smartphones, enormous volumes of data is being generated. Apart from this, a lot of businesses started collecting real time stream data which is voluminous in nature.

Velocity: Data is generated at a faster rate, which makes it even more arduous to manage.

Veracity: This is about the trustworthiness of data.

Variety: Different formats in which data is generated, structured, semi structured and unstructured. A total of 90 percent of the generated in the last two years is unstructured (Sivarajah, Uthayasankar, et al., 2017) [3].

2. Related work

[Manpreet Singh et.al 2017] This author explained how real-time data may be useful to analyze and predict severe emergency cases pretty earlier. An enormous amount of data is generated daily by medical organizations, which collectively include patients, health centers, medical specialists and, of course, diseases. The data are enormous and provide information on future forecasts, which could certainly prevent the occurrence of maximum medical cases. But without the big data analysis techniques and the Hadoop cluster, these data are useless.

[M. D. Anto Praveena & B. Bharathi 2017] Traditional data storage techniques cannot store and analyze these huge amounts of data. Many researchers are doing their research on reducing the size of big data for an analytical and efficient data visualization report. To this the author provided analysis of Big Data, problems, challenges and various technologies related to Big Data.

[Li Zhu et.al 2018] This author discussed some open challenges related to the use of Big Data analysis in Intelligent Transportation Systems (ITS). What can you see in many projects around the world? Intelligent transport systems will produce a large amount of data. The large amount of data produced will have a profound impact on the design and application of intelligent transport systems, making ITS safer, more efficient and more economical. The study of Big Data analysis in the IST is a growing sector.

[Sohail Jabbar et.al 2018] In this author proposed a merging of three different data models, such as relational, semantic, and large data and metadata involving their improved problems and capabilities. The traditional analytics approaches of big-data use the grouping of data in small segments while providing a distributed computation between several secondary nodes. These approaches pose particular problems related to network capacity, specialized tools and applications that cannot be formed over a short period of time.

[Rua-Huan Tsaih et.al 2018] The author analyzed currency information; exchange rate is the value of the currency of one country

Copyright © 2018 R. Venkateswara Reddy, Dr. D. Murali. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

over another. Most major economic organizations have adopted floating exchange rates. In a floating exchange rate system, market forces of supply and demand of foreign and domestic currencies determine the exchange rate. Therefore, in order to stabilize an exchange rate, the government of a country must have substantial reserves to control the supply and demand of foreign and domestic currency.

3. Hadoop distributed file system (HDFS)

HDFS is one of the two main components of the Hadoop framework. Manage the storage area of Hadoop and it is made according to the Google file system (GFS). It is a distributed fault-tolerant file system and can be run on basic hardware (Beakta, Rahul, 2015) [4].
Pig Latin
Pig Latin developed by Yahoo Research and is presently managed by Apache Software Foundation. Pig makes implementing MapReduce jobs relatively easy because of this it is becoming popular data flow programming language. Pigs works on top of HDFS when executed, the Pig statements are converted into MapReduce jobs (Olston, Christopher, et al., 2008) [5].

4. Methodology

The figure 1 represents the workflow of the methodology employed. The first step involves finding or defining the problem statement. In the second phase required data is collected and is copied onto the Hadoop Distributed File System (HDFS) in the third step. In the subsequent step we have performed data cleansing, which is one of the greater significant phases in the analytics. It involves removing duplicate value, fixing the missing values and schema issues. Hadoop clustered environment is used to store and process the cleansed data (Jin, Xiaolong et al, 2015) [6]. HDFS is the file system used in Hadoop environment, which is developed to work on distributed environment. HDFS is horizontally scalable and more reliable than the existing systems and is greatly favored in Big Data processing. HDFS replicated data on three nodes by default and this number can be changed using the below command.
hadoop fs -setrep [-R] [-w] <noOfReplicas><path> ...

- path represents the path of the file or directory
- R option for backward compatibility
- w option for whether the command should wait for the replication to be completed

Alternatively, this can be changed manually from hdfs-site.xml file by changing the dfs.replicationproperty.
Apart from HDFS, the other core component of Hadoop is MapReduce. MapReduce manages the processing of the Big Data. The storing and processing of the data is performed by the two components from MapReduce, Job Tracker and Task Tracker. Job Tracker credits MapReduce activities to task followers and activities that send the heart rate signal to update Job Tracker in the status of the activity assigned to them. (Aghbari, Zaher Al, 2015) [7]. Pig Latin scripting language is used here, with the Pig Latin the implementation time is reduced drastically when compared to the MapReduce.The end results are presented and analysed in graphs.

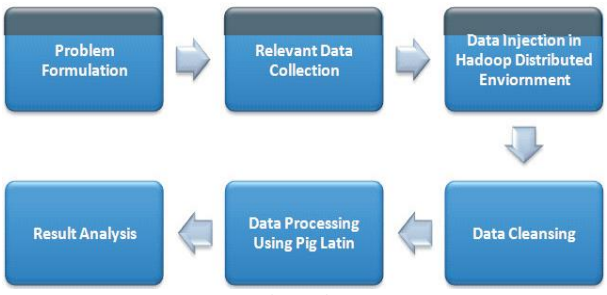


Fig. 1: Workflow of Methodology.

Framing Queries

In India, the health care services and their standards have improved drastically in the last two decades due to the hike in the number of private and government hospitals and upsurge in the number of medical professionals. In this paper, healthcare data of last few decades is evaluated using disparate experimentational queries.
Pig Latin query steps
Finding the total number of hospitals in each year group

- Enter the Pig shell using 'pig' command
- X = Load dataset;
- Y = Sort and group the hospitals;
- Z = Get the total of hospitals in each group;
- Dump Z;

5. Discussions

The figure 2 represents the exponential growth of number of hospitals from 1950 to 2015. This demonstrates that the accessibility of medical services was increasing consistently throughout the mentioned time period.

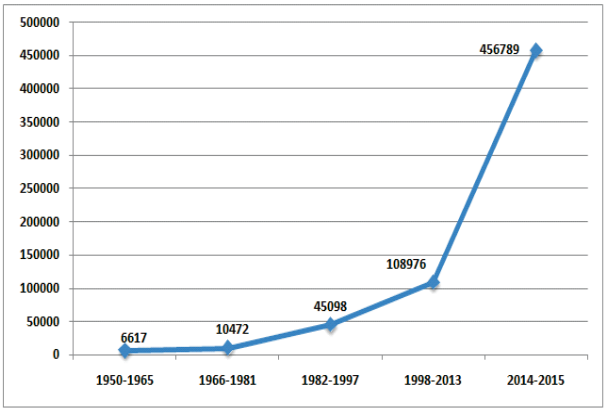


Fig. 2: Number of Hospitals from 1950 to 2015.

Finding number of doctors each year from 2005 to 2015

- Enter the Pig shell using 'pig' command
- X = Load dataset;
- Y = Sort and group the doctors by year;
- Z = Get the total of doctors in each group;
- Dump Z;

The figure 3 represents how the number of doctors were increasing throughout the period of 2005 to 2015.

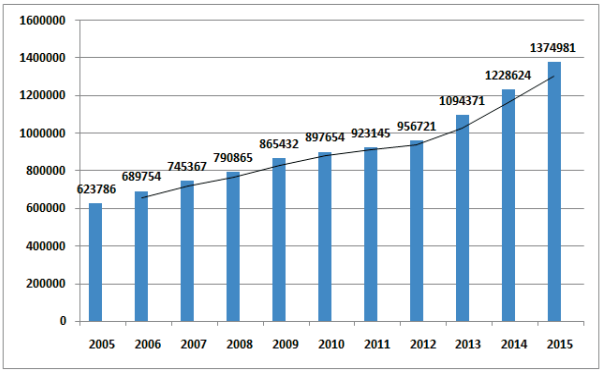


Fig. 3: Number of Doctors from 2005 to 2015.

Finding Life Expectancy Male vs Female

- Enter the Pig shell using 'pig' command
- X = Load dataset;
- Y = Group and Generate male and female life expectancy by state;
- Dump Y;

The figure 4 depicts the male and female life expectancy by each state. It is understood from the graph that Madhya Pradesh has the

least life expectancy for both male and female and Kerala has the highest.

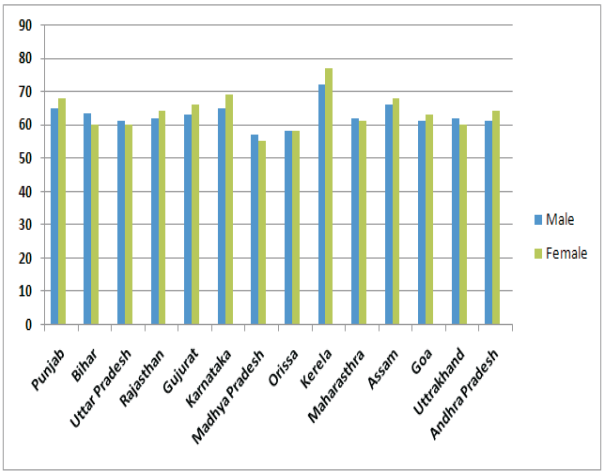


Fig. 4: Male vs Female Life Expectancy Across States.

	Male	Female
Punjab	65	68
Bihar	63.2	60.1
Uttar Pradesh	61	60
Rajasthan	62	64
Gujurat	63	66
Karnataka	65	69
Madhya Pradesh	57	55
Orissa	58	58
Kerela	72	77
Maharasthra	62	61
Assam	66	68
Goa	61	63
Uttrakhand	62	60
Andhra Pradesh	61	64

Fig. 5: Male vs Female Life Expectancy across States.

Finding number of people satisfied with healthcare services

- Enter the Pig shell using 'pig' command
- X = Load dataset
- Y = Sort and group number of people satisfied with medical services by state
- Z = Get the percentage of the satisfied people by state
- Dump Z;

The figure 6interprets the number of people satisfied with medical services in their state. It is pointed out that people from Uttar Pradesh are least satisfied while people from Kerala are most satisfied.

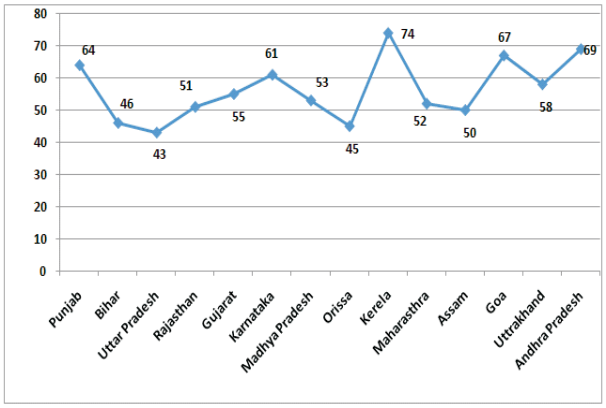


Fig. 6: Percentage of People Satisfied with Healthcare Standards.

6. Conclusion

In this paper, Indian healthcare dataset of the time 1950 to 2015 is analyzed with different research queries using Pig Latin. The healthcare standards in India have boostedconsistently in the last few decades because of the increase in the number of government and private hospitals. The increase in the number of hospitals and doctors been illustrated in the graphs and tables. The number of people satisfied with the healthcare standards in their state is analyzed. This work is to encourage more research into the healthcare and pave way for better standards in medical services.

References

[1] "More CMOs Feeling Unprepared For the 'Data Explosion.'" Marketing Charts, 4 July 2017, www.marketingcharts.com/digital-37207.

[2] Katal, Avita, et al. "Big Data: Issues, Challenges, Tools and Good Practices." 2013 Sixth International Conference on Contemporary Computing (IC3), 2013, doi:10.1109/ic3.2013.6612229.

[3] Sivarajah, Uthayasankar, et al. "Critical Analysis of Big Data Challenges and Analytical Methods." Journal of Business Research, vol. 70, 2017, pp. 263–286., doi:10.1016/j.jbusres.2016.08.001.

[4] Beakta, Rahul. (2015). Big Data and Hadoop: A Review Paper. International journal of computer science & information te. 2.

[5] Olston, Christopher, et al. "Pig Latin" International Conference on Management of Data SIGMOD, 2008, doi:10.1145/1376616.1376726.

[6] Jin, Xiaolonget al., (2015). Significance and Challenges of Big Data Research.

[7] Al Aghbari, Zaher. (2015). Mining Big Data - Challenges and Opportunities. ICEIS 2015 - 17th International Conference on Enterprise Information Systems, Proceedings. 1. 379-384. 10.5220/0005463803790384.

[8] Manpreet Singh et.al (2017), "Big data analytics: Solution to healthcare", 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), pp: 239 – 241.

[9] M. D. Anto Praveena & B. Bharathi (2017), "A survey paper on big data analytics", 2017 International Conference on Information Communication and Embedded Systems (ICICES), and pp: 1-9.

[10] Li Zhu et.al (2018), "Big Data Analytics in Intelligent Transportation Systems: A Survey", ISSN: 1524-9050, pp: 1-16.

[11] Sohail Jabbar et.al (2018), "A Methodology of Real-Time Data Fusion for Localized Big Data Analytics", ISSN: 2169-3536, Volume 6, pp: 24510 – 24520.

[12] Rua-Huan Tsaih et.al (2018), "The use of big data analytics to predict the foreign exchange rate based on public media: A machine-learning experiment", ISSN: 1520-9202, Volume 20, issue 2, pp: 34-41.