International Journal of Research Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

# Big Data Analytics on Indian Healthcare Data

[1]Y.PRASANTHI, [2]K ARCHANA

**ABSTRACT**: Big Data is the tremendous volumes of data, continuing produced at present times. Companies are using this Big Data to examine and prognosticate the future to earn profits and gain competing for an edge in the market. Big Data analytics has implied accepted into almost every field, retail, banking, governance and health care. Big Data can happen appropriated for investigating healthcare data for better planning and better decision making which lead to improved healthcare standards. In this paper, Indian healthcare data from 1950 to 2017 is analyzed using different research queries. This healthcare generates the significant amount of heterogeneous data. But without proper data analytics methods, these data became useless. Big Data Analytics using Hadoop plays an active role in performing meaningful real-time analysis on the enormous volume of data and able to predict the emergency situations before it happens. It describes the big data use cases in healthcare and government.

**Keywords:** Big Data, Healthcare, Hadoop, Pig Latin, HDFS, MapReduce

## INTRODUCTION:

Data is a critical resource which exists in many forms. Big data do not have a standard definition while it denotes in several ways. The term big data is referred to describe the exponential growth of the data flow in various sectors which is too large to process using the available traditional database and software techniques. Often big data is considered to be scary, yet it is an explosion in the field of knowledge. It helps to implement various analytics, which can create an impression on the economic growth, generating events, increasing efficiency over other methods. This critical collection of data often described as three-dimensional namely Volume, Velocity and Variety where some even define with Veracity (Katal, Avita, et al, 2015) [1]..

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) [2].

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

To deal with these characteristics is an expensive turnout. Apache Hadoop is the knight in the shining armour here. Apache Hadoop is a platform that is responsible for providing with the solutions that are cheaper. Hadoop's major components include a distributed file system which is called the HDFS (Hadoop Distributed File System) and a layer for implementation of the processing paradigm-MapReduce.6 Hadoop is an open source system. Hadoop uses a cluster of commodity machines as its nodes, forming a network which is used as a single, logical, storage and computational platform among multiple users or groups.3 Hadoop's MapReduce performs tasks in parallel, automatically and in a synchronized fashion.

**2. RESEARCH METHODOLOGY** The following figure describes the research methodology adopted. The first step involves identification of problem and then adopting appropriate approach and strategy to solve the problem. According to proposed framework first relevant data was collected then the data from the files will be loaded onto the HDFS using copy command. Data cleansing is one of the most important component required in data analytics. It includes data cleansing, data extraction, removing duplicates and converting it into standardized schemas. Storing and processing is done used HDFS.12 HDFS being reliable and scalable is highly recommended for storing and processing large amount of data. HDFS replicates data over multiple nodes and thus does not require RAID (Redundant Array of Integrated Devices) storage.4 The task of storing, accessing and modifying data is performed using two different components Job Tracker and Task Tracker. The Job Tracker assigns the MapReduce tasks to Task Trackers. The Task Trackers send their status of being active and ready to take up the job by sending heartbeats to the Job Tracker.7 The analytical tool here is the scripting language; Pig Latin. Overall development time and testing time is much less than that of map reduce program. Writing fewer codes without having prior knowledge of JAVA and reduced testing time are major advantages of Pig latin over MapReduce program. At the end results were analyzed using graphs to make useful decisions.


Fig. 1. Research Methodology.

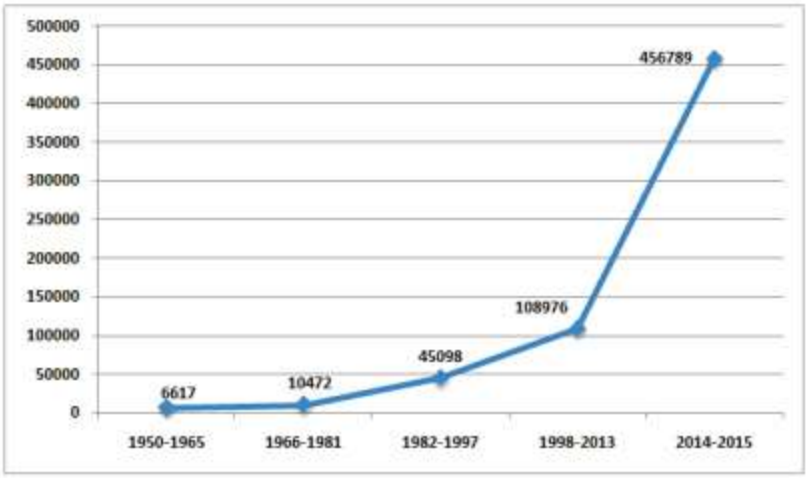


Fig. 2. Aggregate Number of Hospitals from 1950–2015.

**3.QUERY FORMULATION** In this paper, we authors analyzed the health care dataset against different research queries, over the last few decades; quality of health care services in India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized medical qualification. The main objective is to provide the healthcare services to all and to improve the accessibility of healthcare amenities to all the stratum of the society. The data set was in .csv format which was first cleaned by removing the missing values and other inconsistent values. Fully distributed hadoop cluster was used to store and process the voluminous data.

## 4. EMPIRICAL OBSERVATIONS

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

### 4.1 Aggregate number of hospitals from 1950–2015

Input: health care dataset Output: Aggregate number of hospitals from 1950–2015

Creating Pig Script

1. Using Pig command enter into Pig's interactive shell

2. A = Load the healthcare data set;

3. B = Sort and group number of hospitals from 1950–2015;

4. C = Find out the sum of hospitals in each group year;

5. Dump C; Above graph clearly depicts that there is tremendous growth of hospitals from the year 1950–2015. This indicates that the ease of availability of healthcare facilities is increasing every year. In spite of significant growth government has to take strict measures to improve the overall health care facilities in India because considerable amount of gaps do exist between the demands to that of quality supply of healthcare services.

**4.2 Aggregate number of physician from 2005–2015 Input:** Health Care dataset Output: Aggregate number of physician from 2005–2015

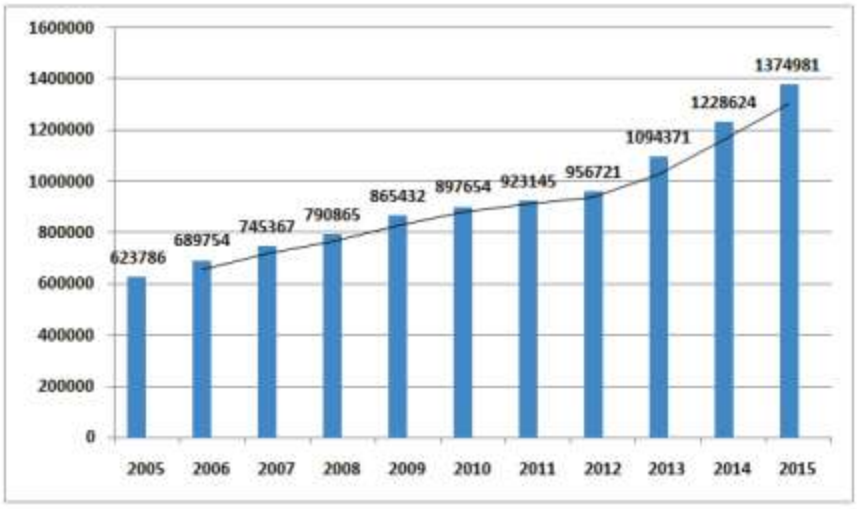


Fig. 3. Aggregate Number of Physician from 2005–2015.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

Creating Pig Script

1. Using Pig command enter into Pig's interactive shell

2. A = Load the healthcare data set;

3. B = Sort and group number of physician from 2005–2015;

4. C = Find out the sum of physician in each group year;

5. Dump C; Above graph highlights that more numbers of doctors with recognized medical qualification have been registered from the year 2005 to 2015. India's doctor – patient ratio is still alarming despite of the fact that there has been tremendous growth in the number of doctors and nurses.

**4.3 Male-female life expectancy in various states Input**: Health care dataset Output: Male-female Life Expectancy in Various states Creating Pig Script

1. Using Pig command enter into Pig's interactive shell

2. A = Load the healthcare dataset;

3. B = Group and generate the life expectancy of male and female in each state;

4. Dump B; From the above graph it could be concluded that the highest male and female expectancy is in Kerala while least in Madhya Pradesh. Therefore government has to come up with different measures to improve the overall healthcare facilities in different state to match up with the international standards of health.

**4.4 Percentage of people happy with the healthcare standards in different state**

Input: Health care dataset Output: Percentage of people happy with the healthcare standards in different state

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

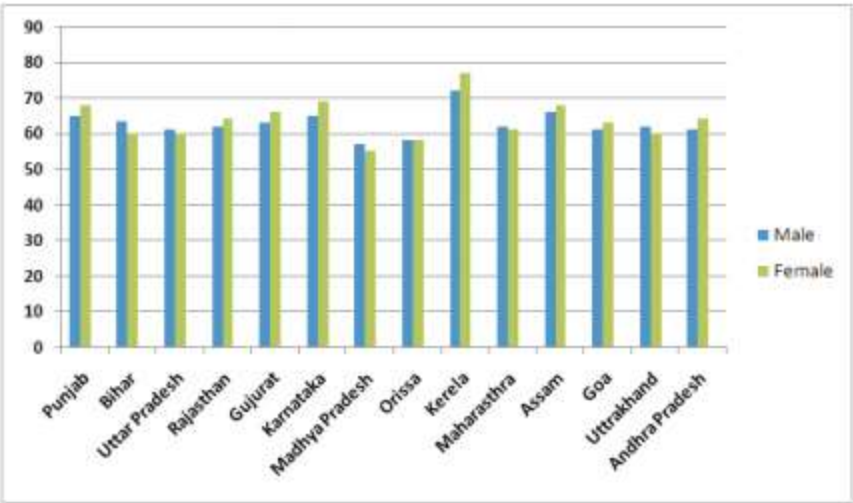


Fig. 4. Male-Female Life Expectancy in Various States.

| | 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-Female Life Expectancy in Various States.

Creating Pig Script

1. Using Pig command enter into Pig's interactive shell

2. A = Load the healthcare data set;

3. B = Sort and group number of happy people with the healthcare standards in each state;

4. C = Find out the percentage of happy people in each state;

5. Dump C; From the above graph it could be highlighted that citizen of Kerala are most satisfied with the healthcare facilities provided in their states while least in Uttar Pradesh. This indicates that the ease of availability of healthcare facilities is more in Kerala as compare to Uttar Pradesh. Government must take strict measures to address all the issues to achieve the objective: To provide the healthcare services to all.

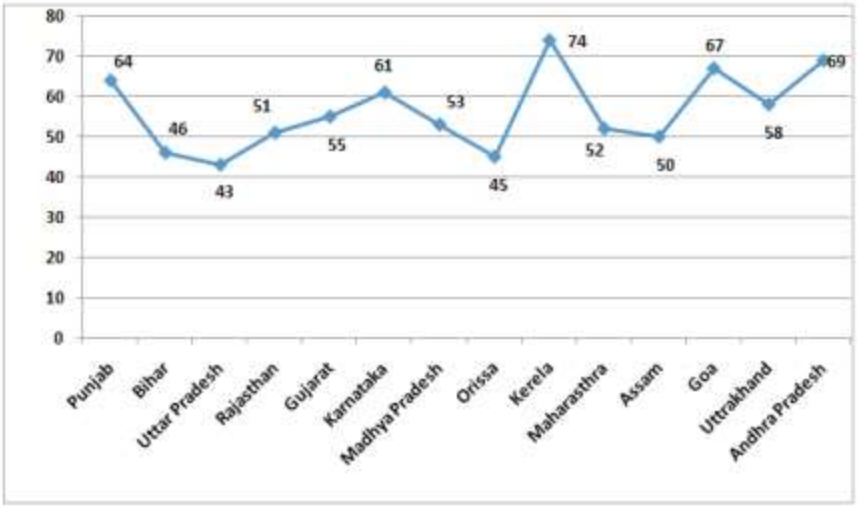


Fig. 6. Percentage of People Happy with the Healthcare Standards in Different State.

**5. LIMITATION** Following are the main limitations:

1. Apart from using Pig latin, MapReduce programming framework can be used to analyze the health care data set.

2. Pseudo distributed hadoop mode us used to implement the data set.

3. The main limitation includes few publications and Non English publications were excluded therefore as authors we could not claim that the work has not been printed in other languages.

**6. CONCLUSIONS** In this paper, we authors analyzed the health care dataset against different research queries using Pig Latin Script, over the last few decades; quality of health care services in India has been improved tremendously because of the improved health care services, increased number of private and government hospitals and increased number of doctors with recognized

medical qualification. The main objective is to provide the healthcare services to all and to improve the accessibility of healthcare amenities to all the stratum of the society. In spite of significant growth government has to take strict measures to improve the overall health care facilities in India because considerable amount of gaps do exist between the demands to that of quality supply of healthcare services

## REFERENCES

[1] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and Taha, K., Efficient Machine Learning for Big Data: A Review, Big Data Research, vol. 2(3), pp. 87–93, (2015).

[2] K. N. Aye and T. Thein, A Platform for Big Data Analytics on Distributed Scale-Out Storage System, International Journal of Big Data Intelligence, vol. 2(2), pp. 127–141, (2015).

[3] A. T. Azar and A. E. Hassanien, Dimensionality Reduction of Medical Big Data Using Neural-Fuzzy Classifier, Soft Computing, vol. 19(4), pp. 1115–1127, (2014).

[4] M. Chen, S. Mao and Y. Liu, Big Data: A Survey, Springer- Mobile Networks and Applications, vol. 19(2), pp. 171–209, (2009).

[5] M. Fedoryszak, D. Tkaczyk and L. Bolikowski, Large Scale Citation Matching Using Apache Hadoop, Springer- Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science, vol. 8092, pp. 362–365, (2013).

[6] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S. U. Khan, The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues, Information System, vol. 47, pp. 98–115, (2015).

[7] A. E. Hassanien, A. T. Azar, V. Snasel, J. Kacprzyk and J. H. Abawajy, Big Data in Complex Systems: Challenges and Opportunities, Verlag GmbH Berlin/Heidelberg: Studies in Big Data, Springer, vol. 9, (2015).

[8] T. Huang, L. Lan, X. Fang, P. An, J. Min and F. Wang, Promises and Challenges of Big Data Computing in Health Science, Big Data Research, vol. 2(1), pp. 2–11, (2015).

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 04
February 2018

[9] S. Ibrahim, H. Jin, L. Lu, L. Qi, S. Wu and X. Shi, Evaluating Map Reduce on Virtual Machines: The Hadoop Case, Springer: Cloud Computing Lecture Notes in Computer Science, vol. 5931, pp. 519–528, (2009).

[10] A. Jacobs, The Pathologies of Big Data, Communications of the ACM – A Blind Person's Interaction with Technology, vol. 52(8), pp. 36-44, (2009).

[11] H. V. Jagadish, Big Data and Science: Myths and Reality, Big Data Research, vol. 2(2), pp. 49–52, (2015).

[12] X. Jin, B. W. Wah, X. Cheng and Y. Wang, Significance and Challenges of Big Data Research, Big Data Research, vol. 2(2), pp. 59–64, (2015).

[13] K. Kolomvatsos, C. Anagnostopoulos and S. Hadjiefthymiades, An Efficient Time Optimized Scheme for Progressive Analytics in Big Data, Big Data Research, vol. 2(4), pp. 155–165 (2015).