Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services

D. Peter Augustine
Assistant Professor
Christ University
Bangalore 560029

ABSTRACT
In this paper, we analyze and reveal the benefits of Big Data Analytics and Hadoop in the applications of Healthcare where the data flow to and from is in massive volume. The developing countries like India with huge population faces various problems in the field of healthcare with respect to the expenses, meeting the needs of the economically deprived people, access to the hospitals, research in the field of medicine and especially in the time of spreading epidemics. This paper gives the involvement of Big Data Analytics and Hadoop and reveals the impact of the same to render the services of healthcare to everyone in the optimal cost.

General Terms
Health care, Big Data Analytics, Hadoop, Hadoop Distributed File System, Map Reducing, Health Care Applications.

Keywords
Health care in India, Big Data Analytics in health care, Hadoop in health care.

1. INTRODUCTION
The exponential growth of data over the last decade has introduced a new domain in the field of information technology called Big Data. Datasets that stretches the limits of traditional data processing and storage systems is often referred to as Big Data. The need to process and analyze such massive datasets has introduced a new form of data analytics called Big Data Analytics. It includes analyzing huge measure of data of a mixture of types to reveal hidden blueprint, unidentified association and other useful information. Many organizations are increasingly using Big Data analytics to get better insights into their businesses, increase their income and profitability and gain competitive advantages over rival organizations.

The distributed processing of outsized data sets across groups of systems is facilitated by using simple computing models of the Apache Hadoop Framework. The framework is proposed to widen from solitary servers to thousands of systems, each presenting native computation and storage. The library itself is designed in such a way that the failures can be detected and handled at the application layer itself. So the framework is capable of yielding all the time service on top of a group of systems prone to failures. This robust future of Hadoop framework attracts variety of companies and organizations to use it for both research and production.

Healthcare is one of the most important areas for developing and developed countries to facilitate the priceless human resource. Nowadays healthcare industry is flooded with enormous amount of data that need validation and accurate analysis. Even though Big Data Analytics and Hadoop can contribute a major role in processing and analyzing the healthcare data in variety of forms to deliver suitable applications, Medical Image Processing is given little more focus in the study and in turn reduces the cost of services to a common man in the country.

2. HEALTHCARE IN INDIA
India takes the second place in the world in its population. Increasing population in India over-burdens the health care structure in the country. Economic scarcity in a large group of people results in poor way in to health care. The main indicators of human development are Longevity, literacy and GDP per capita. Longevity determines the state of health, and is connected with income and education. Longevity can be negatively affected by the weakness in health sector. Human Development Index (HDI of India ranks low (115th) amongst world nations. Limited access to protective and therapeutic health services is one of the foremost reasons for India to face high burden of disease.

Amartya Kumar Sen, an Indian economist and a Nobel laureate has mentioned that "Growth in national income by itself is not enough, if the benefits do not manifest themselves in the form of more food, better access to health and education".

The director of All India Institute of Medical Sciences (AIIMS), Dr MC Misra has pointed out in a talk that "Advances in medical technology and new medicines are indeed a boon, but to work in India they have to be value for money. Most people can't even afford conventional treatments at subsidized prices in public hospitals".

Low on cost, high on quality of care and with a wide range of treatments available — the Indian healthcare system draws over 1.3 million patients from abroad each year. It has generated $3 billion by the end of 2013. A study given in the November, 2013 issue from Harvard Business Review, by authors Vijay Govindarajan and Ravi Ramamurti gave an affirmation to private hospitals in India for "delivering world-class health care, affordably".

Yet, the World Bank data illustrates that 99 percent of India's population cannot manage to pay for these services. Each year, 39 million people are pushed into poverty by spending out of their pocket for healthcare, with households on average give over 5.8 percent of their expenditures to medical care, the data exposes.

In India less than ten percent people have enrolled for some form of health insurance. Fifty eight percent of an average Indian's total annual expenditure is spent out for hospitalizations alone. World Bank data reveals that over forty percent people borrow heavily or sell assets to meet their hospitalization expenses, which forces 39 million people into poverty each year.

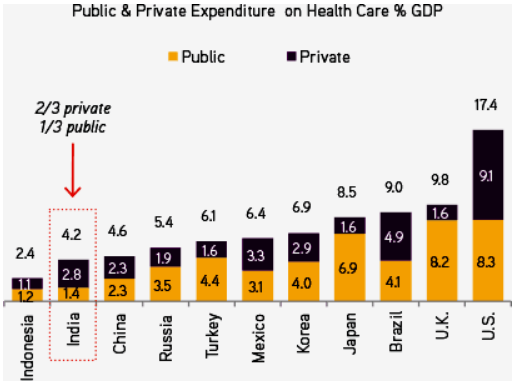


Fig 1: Public and Private Expenditure on Healthcare

IMS Institute of Health Informatics has done a study recent dated 19 July, 2013 has exposed that only 33.33 percent of hospital beds in India can be used by 72 percent of the rural Indian population. But remaining 66.66 percent of total beds are being used by 28 percent of urban Indians. The study also comments that to get the in-patient care in the hospitals, people in the remote villages need to travel at least five kilometers in an average of 63 percent of the time. The statistics from the study is mentioned in the following table.

Table 1. Countries expenditure on healthcare

Country	Total % of GDP spent on Healthcare	Private Expenditure %	Per capita spent on	
			Health care (USD)	Healthcare (USD) by Government
India	**4.1**	**70.8**	**132**	**39**
USA	17.9	46.9	8362	4437
UK	9.6	16.1	3480	2919
South Africa	8.9	55.9	935	412
China	5.1	46.4	379	203
Brazil	9	53	1028	483
Pakistan	2.2	61.5	59	23
Nigeria	5.1	62.1	121	46
Russia	5.1	37.9	998	620

From the statistics one can find that the availability and accessibility of the healthcare to a common man in the society especially developing country like India.

Availability does not mean only the availability of hospitals and healthcare professionals, but it includes the necessary and accurate data on the desktop. Accessibility does not only mean that ability to reach the hospitals but information related to the patient, medicine and all relevant information.

India has a long way to go to reap the benefits by incorporating the information technology effectively for healthcare for its 1.2 billion populations. So it is right time for the public and private sectors in the country to make the technology mandatory for improving the health scenario in healthcare in India.

3. INFLOWING DATA FROM HEALTH MONITORING DEVICES

Despite the government has promised to introduce digitization for maintaining medical records, the reality is not as expected. The country does not even have standardization in common medical terminologies.

Sushil K Meher, computer faculty from AIIMS said that "Our medical terms and names of medicines differ from location to location. For example, the technical term for heart attack is myocardial infarction. But it is called heart attack or even attacked heart by doctors. Similarly, medicines are prescribed instead of the salts they contain. So we cannot maintain a proper record of medical history of patients," said.

Fig 2: Data Inflow for healthcare

Medical records can be considered as an index of a Health Institution. The back bone of Health information system is the department that keeps up those medical records. Medical Record or health record or medical chart is a systematic documentation of a patient's medical history and care. Sometimes medical records are linked to legal report prepared by Medical Officer in obedience to a demand by an authorized Police officer or a Magistrate, and are chiefly referred to criminal cases. Medical records reveal information about the beginning and progress of a Healthcare Center, retrospective and potential statistical analysis, nature of cases admitted to the hospital etc. Medical Records must be carefully and methodically gathered, conserved and secluded for the benefit of every professional in the healthcare system. Medical Records are not only the database of medical and scientific knowledge and inputs to the Government while planning and allocation of budget for health care system of the country. The need of hour is uniformity in storing Medical Records by various Acts.

AIIMS' computer faculty said at a conference on medical informatics organized that "Without the data, we cannot analyze the health situation in India". The health ministry of India had announced recently that the medical records in public hospitals in Delhi will be digitized, so that the information of every patient will be digitized and be accessed by every hospital.

A report by Grant Thornton India tells that by 2014 the Indian health check device and equipment market is expected to grow to around US$ 5.8 billion and by 2016 US$ 7.8 billion, growing at a Compound Annual Growth Rate of 15.5 percent. This rapid development simultaneously produces huge amount of data beyond human thoughts.

4. ROLE OF BIG DATA AND BIG DATA ANALYTICS IN HEALTHCARE

Health care monitoring systems are generating loosely structured data from different sensors that are connected to the patient over a period of time. And these are large complex system requiring efficient algorithms to process these raw data's and require huge computational power. Big data refers to the data generated from different sensors this includes medical, traffic and social data. Some of the characteristics of big data are volume, velocity and value.

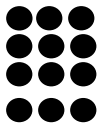
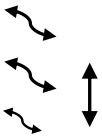
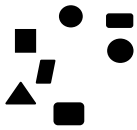
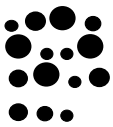
Volume	Velocity	Variety	Veracity
			
Data at Rest TB to XB of data to process	**Data in motion** Streaming data	**Data in many forms** Structured, unstructured	**Data in Doubt** Uncertainty due to data ambiguity

Fig 3: Big Data – 4 "Vs"

4.1 Characteristics of Big Data

4.1.1 Volume

The amount of data generated by the various medical equipment's are larger in size compared to traditional data.

4.1.2 Velocity

The amounts of data stream by medical network are much less compared to the annual data storage capacity of an entire hospital system.

4.1.3 Variety

Traditional data format are less to adapt new type of sensor data types, deployment etc. Whereas nontraditional data such as medical equipment's are easily adaptable to change.

4.1.4 Veracity

It deals with unsure or vague data. There was always the assumption nn traditional data warehouses that the data is certain, clean, and precise. But it is not so in the case of Big Data.

For developing an infrastructure for big data analysis, it requires a mechanism on how to acquire the data, organize these data and process it to extract meaningful information. This can be represented as data acquisition, data organization and data processing.

Data acquisition is one of the major challenges in the big data platforms. Since these systems handle large volume of data, the system requires low latency in capturing the data and using simple query to process larger volume of data.

Since the data's are of larger volume of size, the system needs to take and process the data from the original storage location. Apache Hadoop provides a technology to process these larger volumes of data and at the same time keeping the data on the original data clusters.

In big data processing the data must be process in a distributed environment. The requirement for analyzing data such as medical information requires statistical and mining approach for analyzing the data. Delivering the data in a faster response time will be at higher priority.

The integration of patient data, data on the effects of drug, medical data, Research and Development data and financial records by healthcare and life sciences companies, can help in identifying the patterns that give enhanced and more proactive healthcare. In addition, if healthcare companies can integrate the patient data together with social media content into their data management system, they can even get better collection of information from which the associations concealed inside can be revealed.

The buoyant dream is that healthcare industry will be able to take data from any possible source, strap up appropriate data and investigate it to find answers that facilitates

1. Reductions in cost
2. Reductions in time
3. New research development and optimized findings
4. Smarter diagnosis leads to accurate decision making

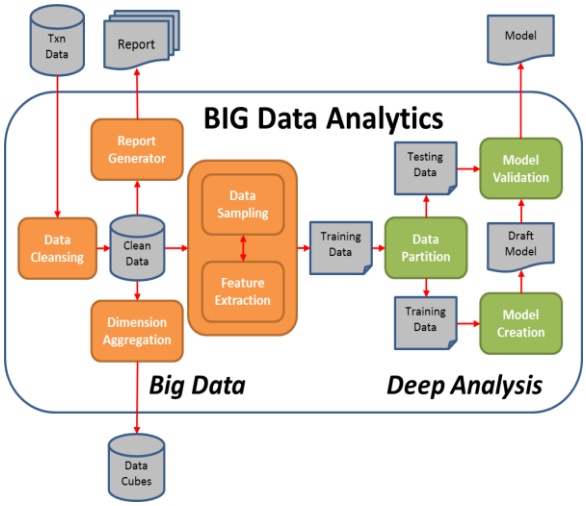


Fig 4: Big Data Analytics

A solution that could meet four key requirements

1. Reliability and scalability in
   a. Storage.
   b. Processing infrastructure.
2. Search engine capabilities for retrieving posts with high availability (HA)
3. Scalable real-time store for retrieving statistics with HA

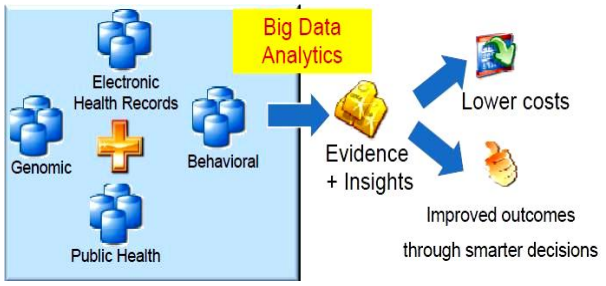


Fig 5: Big Data Analytics in healthcare

In the clinical trial of the future, big data could enable enrollees to not only be monitored for response but also tracked to see if specific subgroups respond differently. Big data approaches are better alternate to traditional clinical trials because they augment the ability to analyze population variability and to conduct analytics in real time. Imagine being able to create dynamic sample size estimations in response to emerging clinical trial data. The key to the value of big data is the ability to enable improvements in clinical trial design that will allow shorter and more efficient trials.

5. HADOOP IN HEALTH CARE DATA

Hadoop has fundamentally changed the economics of storing and analyzing information. As recently as five years ago, a scalable relational database cost $100K per terabyte for a perpetual software license, plus $20K per year for maintenance and support. Today we can store, manage and analyze the same amount of information with a $1,200/year subscription. This difference in economics has attracted a lot of attention and will make Hadoop the centerpiece from which most large-scale data management activities and analyses will either integrate or originate.

Now, with more robust SQL capabilities being coupled to the Hadoop infrastructure and bringing the entire SQL-based ecosystem to the world of Hadoop, the market has expanded by one or two orders of magnitude. No longer is Hadoop just the domain of specialists.

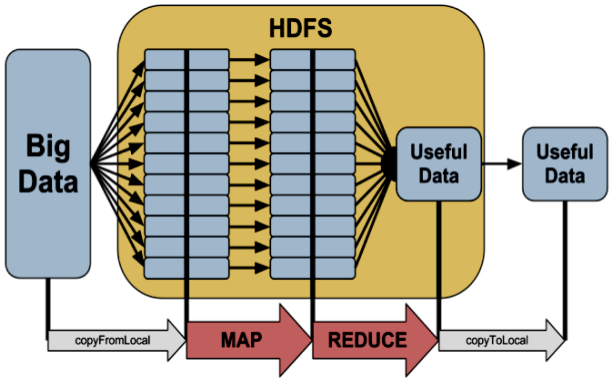


Fig 6: Hadoop's Map Reducing

Relational technologies operating on their own face are another obstacle. The genie is out of the bottle with respect to unstructured/multi-structured/loosely-structured data. With Hadoop we now analyze more and different information than we can with relational databases.

Distributed parallel processing can be facilitated on large volume of data by Hadoop across economical and high level servers can be scaled beyond limits and used for storing and processing the data. No data may be considered for Hadoop as overly big. In the current rapidly growing medical world vast measure of data is being produced and added every day. Because of Hadoop's efficiency and effectiveness the data considered earlier advantageously for the analysis now loses their worth.

Hadoop has the efficiency of storing files on various systems throughout the cluster using a distributed file system. Hadoop hides the location of the files in the cluster being accessed the end users can reference files the same way they do it in the local system.

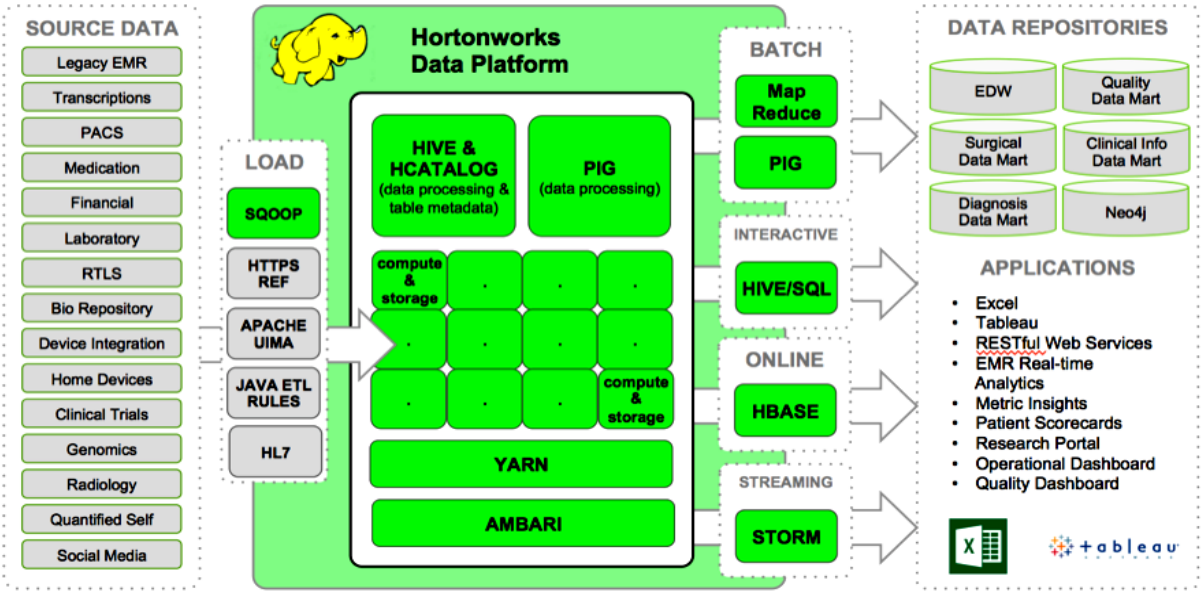


Fig 7: Hortonworks Application using MapReduce

Healthcare field can make significant predictions out of organization through and analyzing Big Data. On the other hand, in view of the fact that 80 percent of medical data is "unstructured", they need to be structured for precise data mining and following analysis. Hadoop is the nucleus proposal for organizing Big Data to give solutions to the difficulty of making Big Data valuable for analytics reasons. Hortonworks Data Platform (HDP) may be considered as potential example to develop the applications for healthcare involving Hadoop and Big Data because of HDP's flexibility, completeness and integration. The following figure shows how Hadoop attempts to map and reduce tasks.

Hadoop tries to run Map and Reduce tasks at the systems where the data being processed is situated when it is doing MapReduce jobs, so that there is no need for data to be copied between systems. The application shows that MapReduce tasks works efficiently when there is one large file as input rather than many small files. Since the small files are residing across many different machines which need significant overheat to copy them to the system where MapReduce takes place. But this is not in the case of large files since they are stored on one system. The application points out that the overhead caused by small files slow the runtime ten to one hundred times. It is evident that the MapReduce framework works well efficiently when the data being processed is localized to the systems doing the process.

There are many challenges faced by the healthcare industry in processing the data to deliver the prolific service to everyone involved in it. One of the areas among them is processing of medical images. Hadoop provides solution to analyze the piling up medical images from various sources and extracts the necessary data to give right diagnosis. The following interface called HIPI explains how image processing can be accomplished in Hadoop.

5.1 HADOOP IMAGE PROCESSING INTERFACE (HIPI)

Hipi gives an API for processing images in the distributed computing environment.

HipiImageBundle (HIB) is the input type given to HIPI. There are set of images put together as one large file with meta data of the images' layout in HIB. A HIB is formed by an already available set of medical images or directly from other sources like medical devices.

A culling function works on the images to check whether they meet the specified criteria and eliminates those who don't adhere with. For example, the images with less than 10 mega pixels can be rejected for analysis. The CullMapper class is then applied on each image that passes the culling test. The Images are given as FloatImage to the Cullmapper class with an associated ImageHeader. A user has the choice of modifying execution parameters explicit to image processing jobs through the HipiJob object while setup. HIPI does not exactly alter any of the default Hadoop MapReduce performance once the Mapper gets into job.

Great care has been taken to ensure that the implementations of the necessary components of Hadoop MapReduce are efficient and effective for image processing tasks.

5.2 Low Cost and greater Analytic Flexibility

Because Hadoop uses industry standard hardware, the cost per terabyte of storage is, on average, 10 times cheaper than a traditional relational data warehouse system. One has to buy a machine, SAN storage and license for the storage in addition to the storage itself. With Hadoop you buy commodity hardware and you're good to go.

In addition to the storage, one can get a bigger bang for his/her buck because it gives the ability to run analytics on the combined compute and storage. The solutions that we had in place previously really didn't allow for that. Even if the costs were equivalent, the benefit we get from storing data on a Hadoop type solution is far greater than what one could get from storing it in a database."

6. FUTURE RESEARCH DIRECTIONS

Since human life involves, healthcare needs deepest analysis to produce precise data for analysis which can never be compromised at any point of time. Even though our analysis on Big Data Analytics and Hadoop for healthcare helps us to optimize the cost and facilities for a last and least man in the country, the following obstacles fall in the pathway to be triumphed over to yield the optimum result.

- From the perspective of non technical medical professionals, it may be difficult to understand and use big data when it is in an unstructured format like text, image, audio or video.

- Second barrier is, in real-time capturing the most important data as it is going on such as major surgeries and convey that to the right people for better analysis.

- A third blockage is storage of the data, and understanding and analysis of data of huge size with the available limited computational facility.

Healthcare industry needs real-time interactivity with data whereas Hadoop can take the use cases into batch processing. Since Hadoop does not hold up indexing and it is not strictly in compliance with ACID model, transactions cannot be managed there.

Image processing is an important key to the future since it will enable the linking images to other types of data for mining. So research on Medical Image Processing takes a lead and it can contribute a lot to this. Research on the advance Cloud storage technology can enable to overcome the data storage problem.

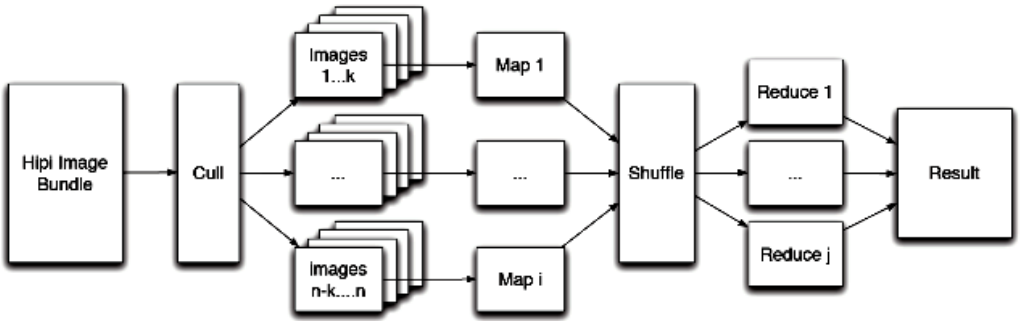


Fig 8: HIPI

7. CONCLUSION

While big data has already been used successfully in consumer markets, challenges remain to its implementation in healthcare. The most important challenges in shift to big data advances are the immense amount of data in the existing systems cannot be related with each other and also the data exist in different file formats. The following challenge for data in the healthcare is to maintain the privacy of the patient while storing and sharing the information interconnected without proper connectivity. It is a burdensome task for institutions that develop applications involving Big Data and Hadoop accordance with acts amended by the National Indian Health Board.

Overcoming these realistic challenges involve the government and its policies, doctors, medical professionals and importantly the technical developers of applications using technology. It is sure that technology itself will help to overcome these underlying problems and contribute to the healthcare industry.

Achieving better outcomes at lower costs has become very important for healthcare, and Big Data Analytics and Hadoop's presence are positively part of the solution in reaching that goal. Although we are in the early days of healthcare big data, it is clear those strategies for big data in the integration of R&D data, efficient clinical trials, and finally in clinical outcomes are foundational to building that solution, lowering costs, and enhances the accessibility and availability of healthcare to all in 1.2 billions of Indians.

8. ACKNOWLEDGMENTS

I thank the expert Dr. Pethuru Raj, Infrastructure Architect from IBM Global Cloud Center of Excellence, Bangalore who has contributed towards development of the template. I would like to thank him for providing various online resources. The immense experience of Dr. Pethuru Raj in developing and implementing Cloud applications, Big Data and Hadoop applications is an eye-opener for me to see how to reap the benefits of Big Data Analytics and Hadoop for the Healthcare especially with the image processing.

9. REFERENCES

[1] http://www.hindustantimes.com/lifestyle/wellness/hospitals-out-of-pocket-and-out-of-reach/. 2013

[2] http://www.deccanherald.com/. 2013

[3] http://freedomhui.com/. 2013

[4] http://www.kkr.com/company/insights/global-macro-trends-13. 2013

[5] http://www.ibef.org/industry/healthcare-india.aspx. 2014

[6] http://www.informationweek.com/big-data/. 2014

[7] Siegel, J. and Perdue, J. Cloud Services Measures for Global Use: The Service Measurement Index (SMI)," *SRII Global Conference (SRII), 2012 Annual* , vol., no., pp.411,415, 24-27 July 2012

[8] Harris, T. 2010. Cloud Computing- An Overview, Whitepaper. Torry Harris Business Solutions.

[9] http://hadoop.apache.org/. 2013

[10] White, T. Hadoop: the Definitive Guide (2nd Edition) [M]. O'Reilly Media, 2010.

[11] Zulkernine, F. Martin, P. Ying Zou. Bauer, M. Gwadry-Sridhar, F. Aboulnaga, A. 2013. Towards Cloud-Based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud, Big Data (BigData Congress), IEEE International Congress.

[12] Yang Song. Alatorre, G. Mandagere, N. Singh, A. 2013. Storage Mining: Where IT Management Meets Big Data Analytics, Big Data (BigData Congress), IEEE International Congress.

[13] Weiyi Shang. Zhen Ming Jiang. Hemmati, H. Adams, B. Hassan, A.E. Martin, P. 2013. Assisting developers of Big Data Analytics Applications when deploying on Hadoop clouds, Software Engineering (ICSE), 35th International Conference.

[14] Mukherjee, A. Datta, J. Jorapur, R. Singhvi, R. Haloi, S. Akram, W. 2012. Shared disk big data analytics with Apache Hadoop, High Performance Computing (HiPC), 19th International Conference.

[15] www.moneycontrol.com/ 2014

[16] Deepak Kumar, B. 2011. "Evaluation of the Medical Records System in an Upcoming Teaching Hospital—A Project for Improvisation", Journal of Medical Systems.

[17] Weiyi Shang. Zhen Ming Jiang. Hemmati, H. Adams, B. Hassan, A.E. Martin, P. 2013. Assisting developers of Big Data Analytics Applications when deploying on Hadoop clouds, Software Engineering (ICSE), 35th International Conference. hipi.cs.virginia.edu/ 2014/ 2014

[18] cs.ucsb.edu/ 2014

[19] Xu Zhengqiao. and Zhao Dewei. 2012. Research on Clustering Algorithm for Massive Data Based on Hadoop Platform, Computer Science & Service System (CSSS), 2012 International.

[20] Hao, Chen. Ying, Qiao. 2011. Research of Cloud Computing Based on the Hadoop Platform, Computational and Information Sciences (ICCIS), 2011 International Conference.

[21] Garcia, T. and Taehyung Wang. 2013. Analysis of Big Data Technologies and Method - Query Large Web Public RDF Datasets on Amazon Cloud Using Hadoop and Open Source Parsers, Semantic Computing (ICSC), IEEE Seventh International Conference.

[22] http://architects.dzone.com/. 2014

[23] Li Xiang. 2011. Analysis on architecture of cloud computing based on Hadoop [J]. Computer Era.