International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 6, December 2015, pp. 1492~1499
ISSN: 2088-8708

De-Identified Personal Health Care System Using Hadoop

Dasari Madhavi, B.V.Ramana
Department of Information Technology, AITAM, Tekkali, A.P.

Article Info

Article history:

Received Jun 12, 2015
Revised Aug 18, 2015
Accepted Sep 4, 2015

Keyword:

Base64
Big Data
Hadoop
Health care records
Map Reduce

ABSTRACT

Hadoop technology plays a vital role in improving the quality of healthcare by delivering right information to right people at right time and reduces its cost and time. Most properly health care functions like admission, discharge, and transfer patient data maintained in Computer based Patient Records (CPR), Personal Health Information (PHI), and Electronic Health Records (EHR). The use of medical Big Data is increasingly popular in health care services and clinical research. The biggest challenges in health care centers are the huge amount of data flows into the systems daily. Crunching this Big Data and de-identifying it in a traditional data mining tools had problems. Therefore to provide solution to the de-identifying personal health information, Map Reduce application uses jar files which contain a combination of MR code and PIG queries. This application also uses advanced mechanism of using UDF (User Data File) which is used to protect the health care dataset. De-identified personal health care system is using Map Reduce, Pig Queries which are needed to be executed on the health care dataset. The application input dataset that contains the information of patients and de-identifies their personal health care. De-identification using Hadoop is also suitable for social and demographic data.

Corresponding Author:

Dasari Madhavi,
Department of Information Technology,
AITAM,
Tekkali, A.P.
Email: dasarimadhavi.it@gmail.com

1. INTRODUCTION

Big Data is a combination of any type of large and complex datasets that it becomes difficult to process using on existing data management tools or traditional data processing applications. Big Data is about real-time analysis and data driven decision-making process.

Big Data is playing crucial role in Health care and several health institutes help to analyzing the large volume of information. Today huge amount of patient data is generated in health care organizations so we can provide the patient care quality and program analysis. By using traditional systems we can't normalize that data because increasing the digitization of health care data means that organizations often add terabytes' worth of patient records to data centers annually.

1.1. Hadoop Innovation in Health Care Intelligence

Many organizations are discovered that their existing data mining and analysis techniques simply not up to yet the task of handling Big Data. One possible to this problem is to build Hadoop cluster. Hadoop is open-source distributed data storage and analysis frame work that access large volume of datasets that may be structured, unstructured and semi structured. Health care data tends to reside in multiple places like EMRs or EHRs, radiology, pharmacy etc.

Aggregating the data which comes from all over the organization into central system such as an Enterprise Data Warehouse (EDW) and make this data accessible and actionable. Hadoop tools in health care industry provide the secure results for analyzing the large volume of patient data at the same time it can give the reliability of clinical outcomes. A successful outcome is a renewal of a prescription in the expected time period. Hadoop can store renewal information and tie it to social media content and online reminders. Hadoop technology can play major role in health care industry, this technology very useful to the public sector; it can improve the patient safety and security.

1.2. Literature Survey

The increasing digitization of healthcare information is analyzing using new techniques for improve the quality of care, health care results, and minimize the costs. Organizations must analyze internal and external patient information to more accurately measure risk and outcomes. At the same time, many clients are working to increase data transparency to produce new insight knowledge.

Praveen Kumar et al [1], in their work proposed that Hadoop is based on Map Reduction is a powerful tool manages the huge amount of data. With this echo system can use fault tolerant techniques. Emad A Mohammed et al [2], in their work big clinical data analytics would emphasize modelling of whole interacting processes in clinical settings and clinical datasets can be evolution of ultra-large-scale datasets. Arantxa Duque Barrachina et al [3] proposed that using Hadoop techniques large datasets can be used to identification of large dataset.

K. Divya et al [4], for protecting the data used a progressive encryption scheme. Hong song Chen [6], in their research article a novel Hadoop-based biosensor Sunspot wireless network architecture, ECC digital signature algorithm, Mysql database and Hadoop HDFS cloud storage; security administrator can use it to protect and manage key data. Lidong Wang et al [7], in their work based on SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, Radio Frequency Identification Technology (RFID).

2. RESEARCH METHOD

For de-identifying dataset following Platforms and tools are used for Big Data analytics in health care. This work follows the procedure as:

1) Data collection
2) Hadoop Cluster and Map reduce
3) Experiments

2.1. Data Collection

In this paper Big Data minimum size consider as a peta byte. Big Data is available in so many sectors like Web and social networking, Machine to machine, Enormous exchange, Biometric sensor data, Human-created data, Gaming industry, Agriculture and Education departments.

2.2. Hadoop Cluster and Map Reduce

Hadoop is a software framework for allows processing of large datasets across the large clusters of computer. Hadoop Distributed File System is a java based distributed file system that can collect all kinds of data without prior organization. Map Reduce is a software programming model for processing large set of data in parallel. Hadoop cluster is interconnected between the HDFS and Map Reduce. So we can implement the program for Hadoop cluster.

2.3. Experiment

Hadoop is an open source framework which is written in java by Apache software foundation and is used to write software application which requires to process huge amount of data. It works in parallel on large clusters which would have thousands of computer nodes on the clusters. It also processes the data very reliable manner and fault-tolerant manner. Hadoop can be installed in cloud era operating system and after completion of Hadoop installation automatically HDFS process will be started with the daemons.

HDFS is having two main layers Master node, Data node. Master Node or Name Node is the master of the system maintained and managed by the Data Node. Master Node can split data into slave node. Data Node or Secondry Name node is provide the actual storage is having the responsibility to read and write for the clients. Map Reduce is an algorithm or concept to process huge amount of data in a quicker way. As per its name it can divide into Mapper and Reducer.

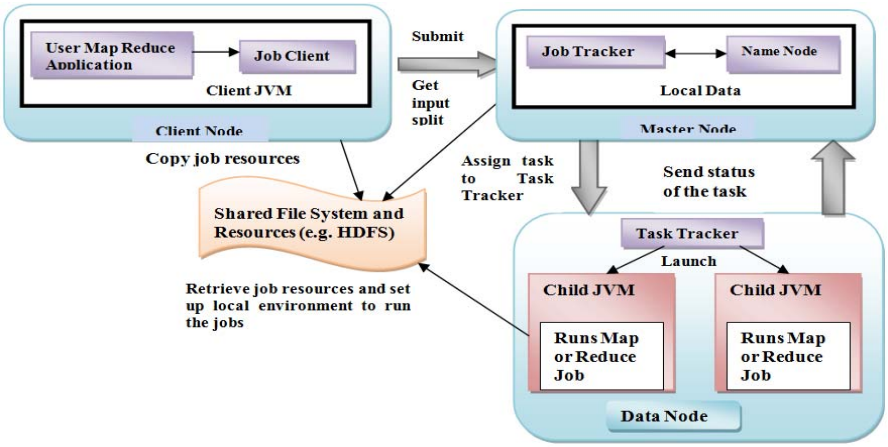

Figure 1. Hadoop application and infrastructure interactions

2.4. Typical Forms of Knowledge De-identification

The HIPAA rule provides protects most individually identifiable health information. There are a couple of common knowledge de-identification tactics that may be deployed to enhance protection in the Hadoop atmosphere, comparable to storage-level encryption and data protecting.

2.4.1. Storage Degree Encryption

Storage-degree encryption the whole quantity that the info set is saved in is encrypted on the disk volume stage while "at relaxation" on the info store, which protects towards unauthorized personnel who can have bodily obtained the disk, from being equipped to read something from it. This is a priceless control in a Hadoop cluster or any massive information store due to common place disk repairs and swap-outs. However this doesn't safeguard the information from access when the disk is running within the system. Decryption technique is applied automatically when the information is read via the running process, and live, inclined data is totally exposed to any user or system gaining access to the method.

2.4.2. Knowledge Protecting

Knowledge overlaying is a priceless manner for obfuscating touchy knowledge, more commonly used for production of test and development knowledge from reside construction know-how. Nonetheless, masked information is meant to be irreversible, which limits its price for a lot of analytic functions and publish-processing necessities. Furthermore there's no warranty that the distinct covering transformation chosen for a detailed sensitive knowledge subject absolutely obfuscates it from identification, mainly when correlated with different knowledge within the Hadoop "data lake" [7] and certain protecting strategies mayor is probably not approved through auditors and assessors, affecting whether or not they real meet regulatory compliance necessities and provide risk less harbor within the event of a data breach.

2.4.3. Cryptography Base 64 algorithm

Cryptography Base 64 is a technique designed to represent an arbitrary sequence of octets (8 bit) in an exceedingly printable text type that enables passing binary information through channels that square measure designed for flat American Standard Code for Information Interchange text like SMTP (Postal, 1982). It additionally permits embedding of binary information in media supporting American Standard Code for Information Interchange text only like XML files. Base64 content Transfer encryption cryptography coding secret writing or Base64 secret writing is outlined in RFC 2045.

3. Results and Discussion

3.1. Preliminary Data Preparation

This work can be involved dummy patient health patient dataset collected in the HSCIC (Health & Social Care Information Centre) contains fields of patient name, patient id, date of birth, Email id, gender, disease, and disease id. That data can be maintained in the format of CSV (Comma Separated Value) file.

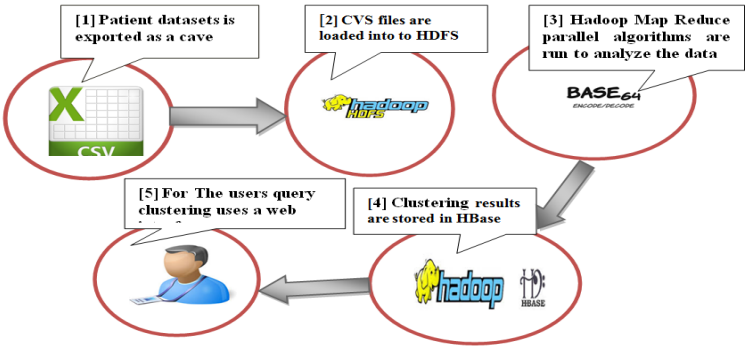

Figure 2. Data Preparation Procedure

3.2. Preliminary Data Analysis

The dataset is in CSV (comma Separated Value) format. The results in this project consisting by using Base 64 encoded algorithm encrypt the plain text into encrypted data. Here using Hadoop single node system then am setting class path for Hadoop jar files.

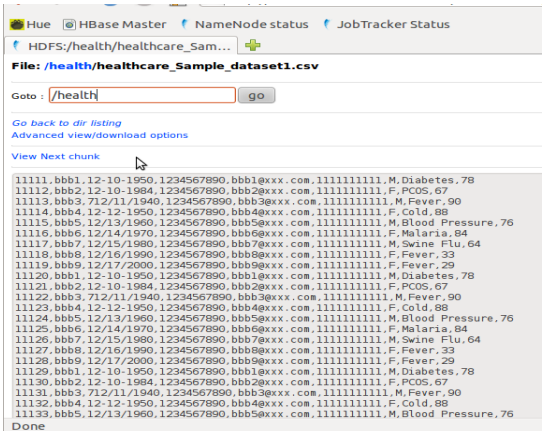

Figure 3. Health care_sample_Dataset1 (plain text)

Export CLASSPATH=.${HADOOP_HOME}/Hadoop-core-0.20.2-cdh3u6.jar :${ CLASSPATH}
Export CLASSPATH=.${HADOOP_HOME}/commons-codec-1.4.jar :${ CLASSPATH}

After the run my java code for mapper, reducer & combiner.

Javac DeIdentifyData.java

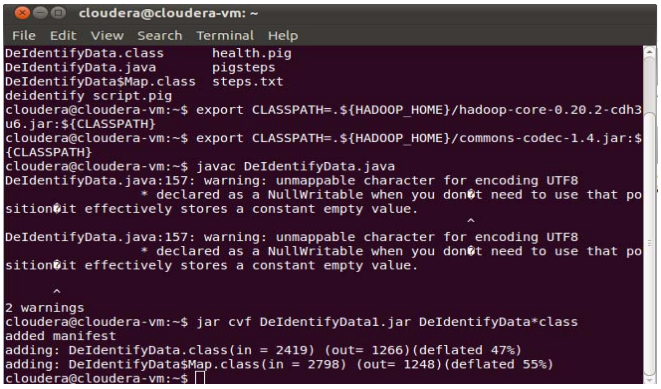

Figure 4. Result of added manifest

3.3. Hadoop Cluster Result

After that placing dataset into Hadoop distributed file system, and run the Hadoop job and finally got the output file.

Hadoop fs -put health care_Sample_dataset1.csv /health/health care_Sample_dataset1.csv
Hadoop fs -put health care_Sample_dataset1.csv /health/health care_Sample_dataset1.csv
Hadoop jar /home/cloudera/DeIdentifyData1.jar DeIdentifyData /health deintout124

Internally the Mapper and Reducer process will be started.

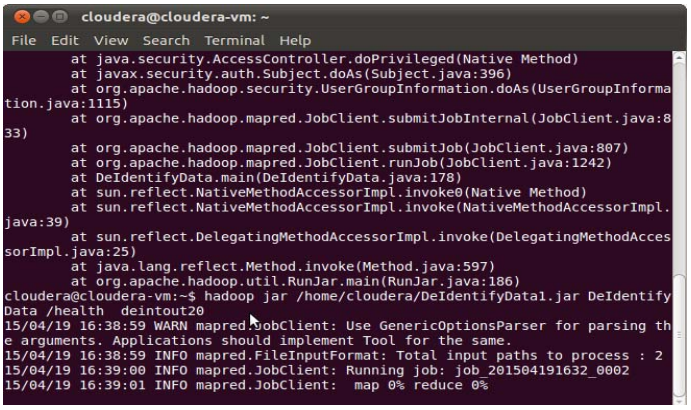

Figure 5.1. Result of Mapper &Reducer initial stages

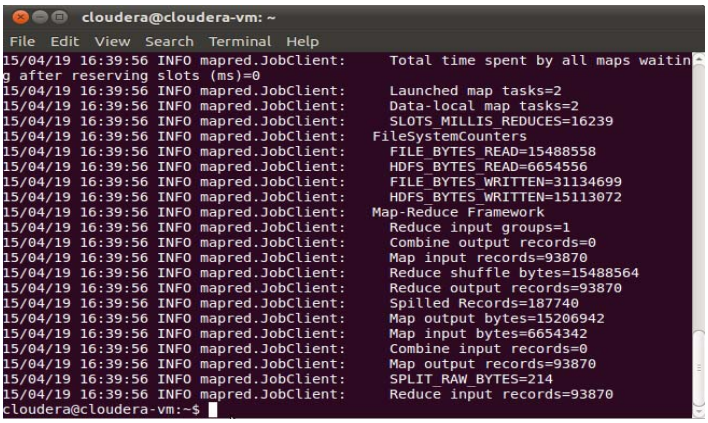

Figure 5.2. Result of status of the Mapper and Reducer

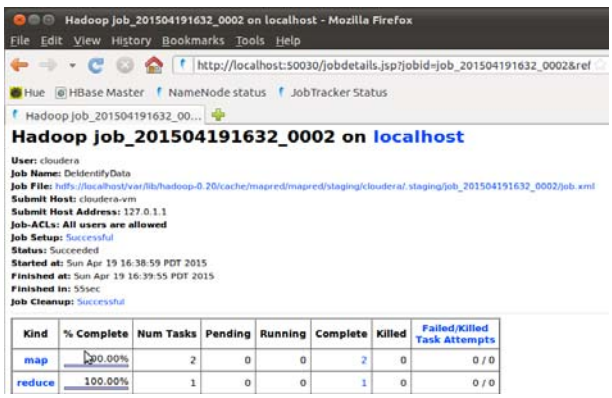

Figure 5.3. Result of Mapper &Reducer Final stage

Then we can check the name node status and job tracker in the browser.

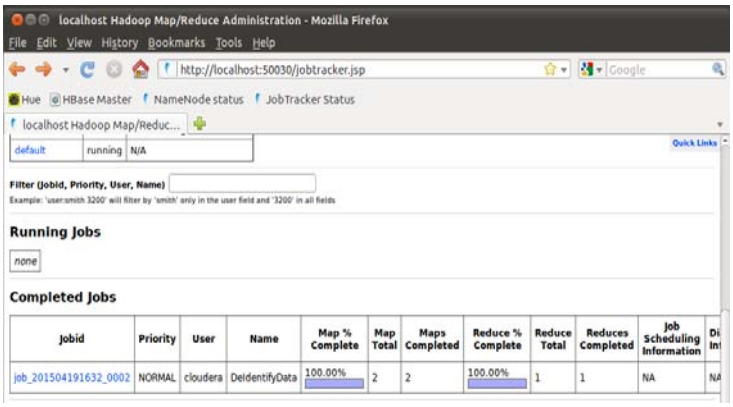

Figure 5.4. Result of the output job file details

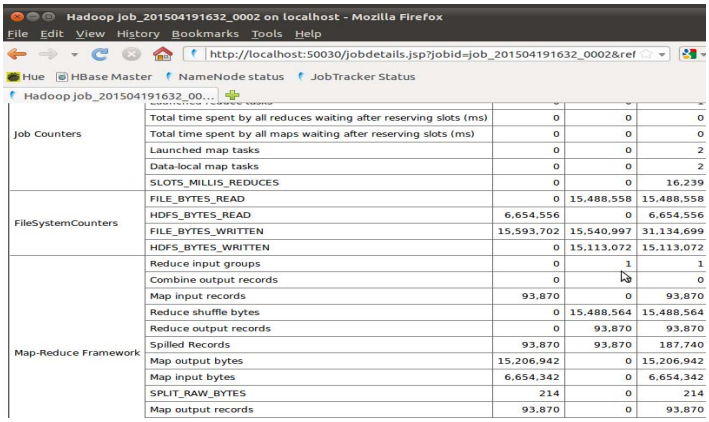

Figure 6. Result of Hadoop Counters

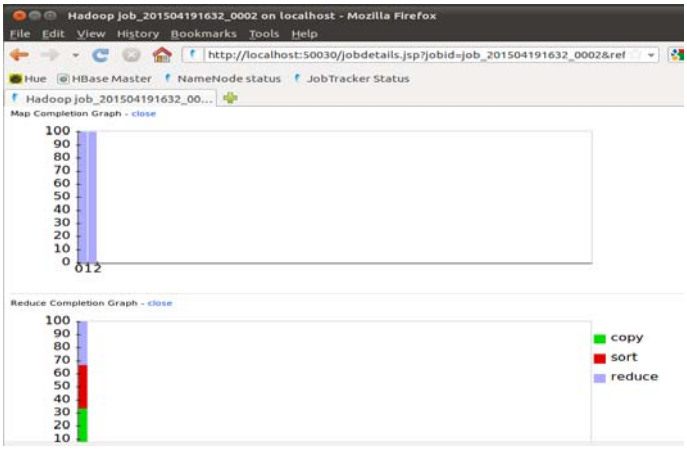

Figure 7. Result of Mapper and Reducer completion graph

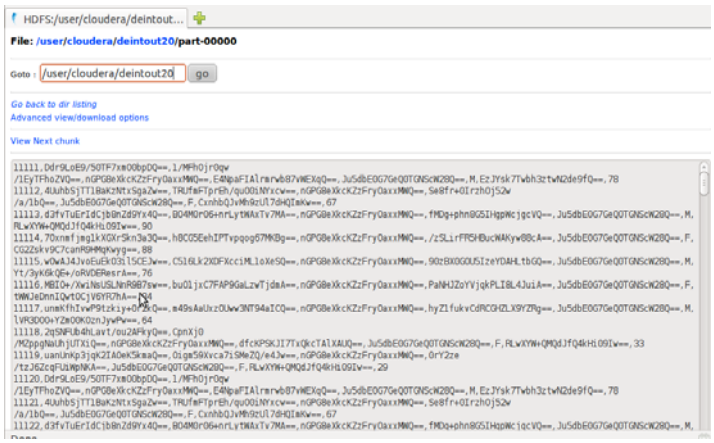

Figure 8. Result of final decrypted output (Encrypted Text)

4. CONCLUSION

Big Data analytics can possibly change the way medicinal services supplier's utilization complex advancements to pick up understanding from their clinical and other information vaults and settle on educated choices. Later on we'll see the fast, across the board execution and utilization of big information examination over the social insurance association and the human services industry. To that end, the few difficulties highlighted above, must be tended to. As large information investigation gets to be more standard, issues ensuring security, defending security, building benchmarks and administration, and without a break enhancing the instruments and innovations will collect consideration. Big Data applications in medicinal services are at an early phase of advancement, yet fast advances in stages and devices can quicken their developing procedure.

REFERENCES

[1] Praveen Kumar, *et al.*, "Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, No. 6, 2014.
[2] Emad A. Mohammed, *et al.*, "Applications of the Map Reduce programming Frame work to clinical Big Data analysis: current landscape and future trends", *Big Data Mining*, 2014.
[3] Arantxa Duque Barrachina and Aisling O'Driscoll, "A Big Data methodology for categorising technical support requests using Hadoop and Mahout", *Journal of Big Data*, 2014.
[4] K. Divya, N. Sadhasivam, "Secure Data Sharing in Cloud Environment Using Multi Authority Attribute Based Encryption", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 1, 2014.
[5] Sabia and Sheetal Kalra, "Applications of Big Data: Current Status and Future Scope", *International Journal of Computer Applications*, Vol. 3, No. 5, pp. 2319-2526, 2014.
[6] Hongsong Chen and Zhongchuan Fu, "Hadoop-Based Healthcare Information System Design and Wireless Security Communication Implementation", *Hindawi Publishing Corporation Mobile Information Systems*, 2015.
[7] Lidong Wang, Cheryl Ann Alexander, "Medical Applications and Healthcare Based on Cloud Computing" *International Journal of Cloud Computing and Services Science (IJ-CLOSER)*, Vol. 2, No. 4, pp. 217-225, 2014.
[8] Priyanka K, Prof. Nagarathna Kulennavar, "A Survey on Big Data Analytics in Health Care", *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 4, pp. 5865-5868, 2014.
[9] Aditi Bansal, Ankita Deshpande, Priyanka Ghare, Seema Dhikale, and Balaji Bodkhe, "Healthcare Data Analysis using Dynamic Slot Allocation in Hadoop," *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 3, No. 5, pp. 2277-3878, 2014.
[10] A Technical Review on, "Protecting Big Data Protection Solutions for the Business Data Lake", *White paper*, 2015.
[11] D. Peter Augustine, "Leveraging Big Data Analytics and Hadoop in Developing India's Health care Services" *International Journal of Computer Applications*, Vol. 89, No. 16, 2014.
[12] Muni Kumar N, Manjula R., *et al.*, "Role of Big Data Analytics in Rural Health Care -A Step Towards Svasth Bharath", *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 6, pp. 7172-7178, 2014.

BIOGRAPHIES OF AUTHORS

D.Madhavi has received B.Tech degree in Information Technology from JNTU, Kakinada. She is currently an M.Tech, Information Technology student in an autonomous institute Aditya Institute of Technology and Management, Tekkali, India. Affiliated to JNTU, Kakinada. Her areas of interests Data Mining and Network Security.

Dr.B.V.Ramana obtained his doctor's degree from Andhra University, India. He is currently working as Professor and Head of the department of Information Technology, Aditya Institute of Technology and Management, India. He has published 18 papers in international Journals and conferences.