Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 50 (2015) 408 – 413

Procedia Computer Science

2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)

A Survey Of Big Data Analytics in Healthcare and Government

J.Archenaa[1] and E.A.Mary Anita[2]
[1]Research Scholar,AMET University,Chennai, [2]S.A.Engineering College,Chennai
E mail: archulect@gmail.com

Abstract

This paper gives an insight of how we can uncover additional value from the data generated by healthcare and government. Large amount of heterogeneous data is generated by these agencies. But without proper data analytics methods these data became useless. Big Data Analytics using Hadoop plays an effective role in performing meaningful real-time analysis on the huge volume of data and able to predict the emergency situations before it happens. It describes about the big data use cases in healthcare and government.

Keywords: Big Data,Hadoop,Healthcare,Map-Reduce

1. Introduction

The healthcare industry has generated large amount of data generated from record keeping, compliance and patient related data. In today's digital world, it is mandatory that these data should be digitized. To improve the quality of healthcare by minimizing the costs, it's necessary that large volume of data generated should be analysed effectively to answer new challenges. Similarly government also generates petabytes of data every day. It requires a technology that helps to perform a real time analysis on the enormous data set. This will help the government to provide value added services to the citizens. Big data analytics helps in discovering valuable decisions by understanding the data patterns and the relationship between them with the help of machine learning algorithms(1). This paper provides an overview of big data analytics in healthcare and government systems. It describes about big data generated by these systems, data characteristics, security issues in handling big data and how big data analytics helps to gain a meaningful insight on these data set.

J.Archenaa. Tel.: 9566156511 *E-mail address:* archulect@gmail.com

Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15)
doi:10.1016/j.procs.2015.04.021

2. Big Data Use Cases

Big data in health-care refers to the patient care data such as physician notes, Lab reports, X-Ray reports, case history, diet regime, list of doctors and nurses in a particular hospital, national health register data, medicine and surgical instruments expiry date identification based on RFID data. Healthcare organizations are depending on big data technology to capture all of these information about a patient to get a more complete view for insight into care coordination and outcomes-based reimbursement models, health management, and patient engagement.

2.1 *Need for Big Data Analytics in Healthcare*

To improve the quality of healthcare by considering the following:

Providing patient centric services: To provide faster relief to the patients by providing evidence based medicine--detecting diseases at the earlier stages based on the clinical data available, minimizing drug doses to avoid side effect and providing efficient medicine based on genetic makeups[1]. This helps in reducing readmission rates thereby reducing cost for the patients.

Detecting spreading diseases earlier: Predicting the viral diseases earlier before spreading based on the live analysis. This can be identified by analysing the social logs of the patients suffering from a disease in a particular geo-location[1]. This helps the healthcare professionals to advise the victims by taking necessary preventive measures.

Monitoring the hospital's quality: Monitoring whether the hospitals are setup according to the norms setup by Indian medical council. This periodical check-up helps government in taking necessary measures against disqualifying hospitals.

Improving the treatment methods: Customized patient treatment---monitoring the effect of medication continuously and based on the analysis dosages of medications can be changed for faster relief. Monitoring patient vital signs to provide proactive care to patients. Making an analysis on the data generated by the patients who already suffered from the same symptoms, helps doctor to provide effective medicines to new patients.

2.2 Need for Big Data in Government

Big data analytics helps government in building smart cities by providing faster and reliable services to its citizens.

Addressing Basic Needs Quickly: Today people need to wait for a long time to get EB, telephone, water, ration card and gas connection. These are the basic needs of citizen. It is the responsibility of the government to provide

these services as quick as possible[3]. Big data analytics plays a major role in achieving it because the data will be analysed on daily basis. People who are in need will be served immediately.

Providing quality education: Education is one of the valuable assets that can be given to the children. It is the duty of government to provide quality education to children[9]. BDA provides detailed report of children who are in the age to be admitted to the school. This helps government to assess the educational needs for these children immediately.

To reduce unemployment rate: To minimize unemployment rate by predicting the job needs before based the literacy rate. This can be achieved by analysis the students graduating each year. It enables government to arrange for special trainings in order to build young entrepreneurs[6].

Other Benefits

- To provide pension to senior citizens without any delay.
- To ensure that benefits provided by government reaches all the people.
- To control traffic in peak times based on the live streaming data about vehicles.
- To monitor the need for mobile ambulance facilities.

3. Implementation of Big Data Ecosystem

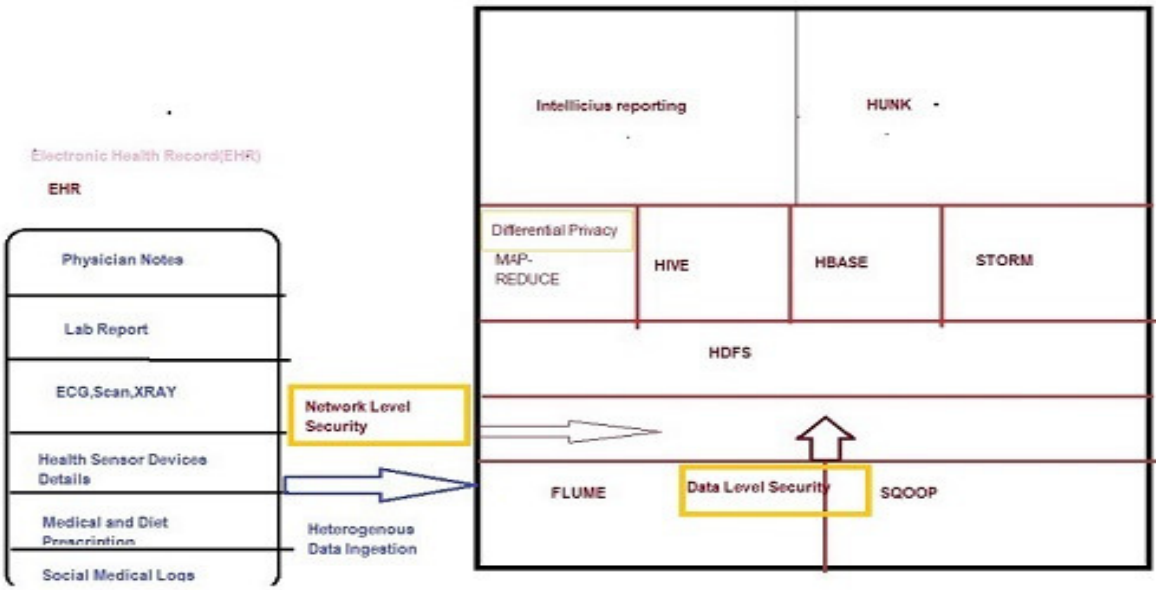

Fig.1. Big Data Healthcare Architecture

The Secured Big Data architecture of healthcare is shown in figure 1. Electronic health record is a heterogeneous data set which is given as input to HDFS through flume and sqoop. Analysis on the data is performed using Map-Reduce and HIVE by implementing machine-learning algorithms which helps in analysing similar pattern of data. This helps in predicting the risk of patient health condition at the earlier stages. Hbase is used for storing the multi-structured data. STORM is used to perform live streaming and any emergency conditions such as patient temperature rate falling beyond the expected level can be intimated to care-takers immediately through AWS Lambda function. Report is generated through intellicius and hunk.

3.1 Big Data Ecosystem for Healthcare and Government

It is a complex system that constitutes of components and technologies to handle large scale data processing and analytics on it. It includes getting the data from various sources, store them in HDFS (Hadoop Distributed File System), process the data using Hadoop components such as Map-Reduce, perform analysis using PIG and generate Business Intelligence reports such as patient scorecards.

3.3 Big Data Lifecycle

3.3.1 Data Collection:It involves the collection of data from various sources and storing it in HDFS. Data can be anything such as case history,medical images, social logs, sensor data etc.

3.3.2 Data Cleaning: It involves the process of verifying whether there is any junk data or any data that has missed values. Such data needs to be removed.

3.3.3 Data Classification: It involves the filtering of data based on their structure. For example Medical Big data consists of mostly unstructured data such as hand written physician notes. Structured, semi-structured and unstructured data should be classified in order to perform meaningful analysis.

3.3.4 Data Modelling: It involves performing analysis on the classified data. For example Government may require the list of malnourished children in a particular location[8]. First it has to classify the data based on the specific location, need to trigger the health report of children, need to identify the children whose family are under poverty line and these data should be processed.

3.3.5 Data Delivery: It involves the generation of report based on the data modelling done. Based on the example after the data is processed it will generate a report based on malnourished children in a particular location. This will help the government to take necessary measures to avoid any further complications.

At the all the stages of BDLC (Big Data Lifecycle) it requires data storage, data integrity and data access control.

3.4 Secured Big Data Architecture

The security challenges faced by the big data processing in distributed environment are as follows:

1. To provide network level security
2. To provide authentication for users, nodes and applications involved in distributed environment.
3. To enable logging in distributed environment for identifying the malicious hackers.

Fig-1 represents secured layered architecture for BDA. Network level security is enforced by using SSL for the communication through RPC between distributed nodes. Two-way authentication can be provided for the data at rest and data in motion. Data stored in database can be encrypted. Data can be transmitted between nodes can encrypting it with Attribute based encryption method. This is effective in preventing the data from malicious users. Inbuilt logging can be implemented in JVM of Map-Reduce using differential privacy to store the user identity where the map-reduce job is done[2]. This helps to identify who is responsible for the leakage of sensitive data.

4. HDFS Architecture

Hadoop effectively handles the large data set. The below figure represents how a client contacts namenode for processing the data. Namenode communicates to Job Tracker and assign the task given by the client for eg to find out the list of patients who are in the risk of getting diabetes. Map reduce program performs the analysis on the data and returns the results to job tracker[4]. It also returns the block where the client can store its data. HiveQL is used to perform the data-warehousing task and it can also be combined with map-reduce program.

PIG provides the platform for analysing large data sets through parallel computations.

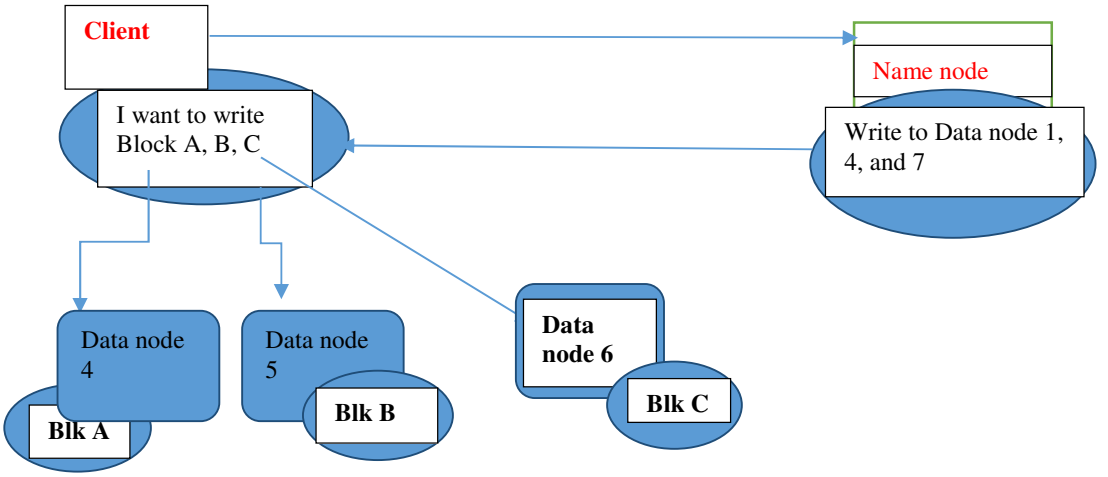

Fig.2. HDFS File System Architecture

The daemons in hdfs are

- ***Name node*** **:** It is the master node which receives the request from the client (example patient monitoring system). It looks up the Meta data to find out which is the suitable data node for storing the data related to the client. It selects data node based on the locality and available free slots.

- ***Secondary Name node:*** It is the backup node for the name node. It stores the fsimage file which contains the details about the data node. Fsimage has to be restored from the secondary name node when name node fails[5].
- ***Job Tracker :***Map reduce program running in job tracker assigns job to the data node and task tracker. Data node stores the actual data and it periodically sends heartbeat to the name node about the data stored. Task tracker performs the task assigned by job tracker.

5. Conclusion

The problem is not the lack of data but the lack of information that can be used to support decision-making, planning and strategy. The entire government system can realize benefits from utilizing big data technologies. To successfully identify and implement big data solutions and benefit from the value that big data can bring, government need to devote time, allocate budget and resources to visioning and planning. With the help of Hadoop the goal of effective citizen care management can be achieved by providing an effective data driven services to citizens by predicting their needs based on the analysis of survey conducted among different classes of citizens. Secured BDA can be implemented by using Hadoop in a security enabled linux environment where access control is provided by the system itself[2].

References

1. Yanglin Ren, *Monitoring patients via a secure and mobile healthcare system, IEEE Symposium on wireless communication,*2011
2. Dai Yuefa, Wu Bo, GuYaqiang ,*Data Security Model for Cloud Computing, International Workshop on Information Security and Application*,2009
3. Jeffrey Dean and Sanjay Ghemawat,*MapReduce: Simplified Data Processing on Large Clusters,*ACM,2008
4. Cong Wang, *Privacy-Preserving Public Auditing for Secure Cloud Storage,IEEE,*2010
5. Konstantin Shvachko, HairongKuang, Sanjay Radia, Robert Chansler ,*The Hadoop Distributed File System,IEEE,*2010
6. Bill Hamilton, *Big Data Is the Future of Healthcare, Cognizant white paper,* 2010.
7. White Paper by SAS, *How Government are using the Power of High Performance Analytics*, 2013.
8. Michael Cooper &Peter Mell, *Tackling Big Data Problems*, NSIT Computer Security Workshop
9. Jean Yean, *Big Data Bigger Opportunities*, White Paper, U.S.General Services Administration.