Review Article

Hadoop Use in Clinical Data Management

Kondaveeti Sreeja*[1], Yanapu Naga Venkata Satya Anusha[1], C.S. Mujeebuddin[2]

1. Intern at Clinosol Research Private Limited, 48-7-53, Rama Talkies Rd, Srinagar, Rama Talkies, Dwaraka Nagar, Visakhapatnam, Andhra Pradesh 530016, India.

2. Founder and CEO of Clinosol Research Private Limited, 48-7-53, Rama Talkies Rd, Srinagar, Rama Talkies, Dwaraka Nagar, Visakhapatnam, Andhra Pradesh 530016, India.

***Corresponding author's E-mail:**

Received: 08-02-2020; **Revised:** 15-03-2020; **Accepted:** 24-03-2020.

ABSTRACT

Big data plays a vital role in health care sector, because it is useful in anticipating the results of disorder eradication of concomitant and death conditions as well as cutting down the amount used for treating illness. Most of the regions in the world are using big data for data storage, it helps for administration and therapy to treat the disorder. Therefore, more number of challenges for developing big data in the field of health care particularly in terms of seclusion, safety, authority, standards, data merging, data arrangement and embodiment of technology. The major target of our project is to establish software that can examine the data with respect to people with different diseases. The actual time scrutiny can be executed by using Hadoop in health analytics for investigating massive volume of data expecting patient's crisis in advance. Paper examination can be a progress of Hadoop, it defines big data gives more possibilities and concerns for health care systems, so it recommends solutions and different technologies which helps the health care organizations to take advantage of this thriving trend.

Keywords: Hadoop, Health care organizations, big data, Health care systems, health analytics, Hadoop software.

INTRODUCTION

HADOOP SOFTWARE which is accessible to the public that allows data storage and operates different applications by using computer hardware. It enhances the trait with foreign data collected from healthcare systems[1,2].

History

In 2002, while working on **APACHE NUTCH PROJECT,** Mike Cafarella and Doug Cutting commenced their work on Hadoop Software. It is between 2003 and 2013 many big data versions were launched and finally in December 2017, Apache Hadoop version 3.0 was implemented and being used as of today. This costs around $30,000 approximately in a month which is very expensive. They recognized that their framework does not work with thousands of data across the network.[3]

Apache Hadoop consists of 2 sub projects-

1. Hadoop MapReduce: MapReduce is a computational model and software framework for writing applications which are run on Hadoop. The MapReduce programs are capable of processing enormous data in parallel on large clusters of computation nodes.

2. HDFS (Hadoop Distributed File System): I t takes part in storage of Hadoop applications. MapReduce consume data from HDFS. It creates a greater number of same data blocks and distributes them on compute nodes in a cluster, this allows reliable and high rapid fluctuations.

Table 1: Development of Hadoop software[4]

In the year	Growth of Hadoop
2003	GFS (Google File System) was released by google to store large information.
2004	MapReduce, which gives resolution for dealing with large collection of data.
2005	Both GFS and MapReduce were commenced by Doug cutting in Apache Nutch Project.
2006	Yahoo was joined along with Nutch project due to some drawbacks in Nutch.
2007	Successfully examined the Hadoop by yahoo and launched Hadoop software.
2008	Hadoop started working with Apache Software Foundation released by yahoo.
2009	After leaving yahoo, Doug cutting affixed with cloudera for expelling Hadoop software with other organizations.

In 2011, Apache Software Foundation launched Apache Hadoop version 1.0.

In 2013, Hadoop version 2.0.6 was available.

Finally, in December 2017 Apache Hadoop version 3.0 was launched and currently in use.[4]

Physicians should understand and harmonize the two major types of data.

- Structured Data.
- Unstructured Data.

Structured Data

- Storage of data within the limits.
- It can easily examine and reserve the information as it has uncomplicated borders.
- Patient related data, Detection of disease; process and medicine codes; remaining information from e-health documents has been produced as categorized in a unique way.
- Data warehouses are used to store the structured data.

Unstructured Data

- This is undefined data which is not suitable to health care sectors it may be contextual and non – contextual.
- It's not only confined to social networking but also to some other forms.
- As the data is unspecified, it cannot be examined similarly as that of structured data.[3]

Table 2: Differences between structured and unstructured data[9]

	Structured data	Unstructured Data
Properties	Easy to find	Requires much effort to search data.
Produced by	Instruments or Individuals	Individuals or Instruments.
Resides in	Data stockroom	NOQSL Databases Data Stockroom
Examples	• Dates • Phone numbers • Credit card numbers • Addresses • Customer names • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Importance - Less expensive, error tolerant, extensibility, processing capability, adaptability and protection of data. It has the ability to operate thousands of terabytes of information or data.[2]

CLINICAL DATA MANAGEMENT

It plays a key role in collecting and managing research information obeying regulatory standards to acquire quality data at the stage of clinical research. It is due to growing demand from both regulatory and pharmaceutical industry, the field of CDM is taken into account.[5]

Clinical data management undergoes 3 steps:

1. Collection.
2. Cleaning.
3. Management of data.

Hadoop's use in clinical data management comes throughout the network for sharing and providing data, where e-health care organizations get sophisticated. The main task for healthcare organizations is to understand whole information and their benefits being used for surgeries, clinical research and course of therapy.[1] Numerous hospitals globally started using Hadoop to assist the hospital teams to perform their tasks effectively with tremendous information. Almost 80% of data present in the hospitals comes under unstructured data, it is impossible to analyse the unstructured data in health care organizations without Hadoop.[6]

Innovative Medical Services

Better Functioning Capacity

Improving Medications

Compilation of data within the country and across the world offers healthcare studies with more number of subjects for clinical trials, coursing, epidemic disease surveillance and screening for better output in order to get notified by physicians before an adverse event occurs[1].

Billions and trillions of unstructured data can be stored. MapReduce engine and HDFS (Hadoop distributed file system) have the capacity to operate thousands of terabytes data can simply identify the patterns in the scope of scam recognition.[6] These costs around $30,000 approximately in a month which is very expensive.[4]

Advantages

- Big data helps to find an intervention before the expected time, so the patients can be treated by giving prophylaxis medication. By this a patient can be rescued from death.
- The aim of Hadoop in clinical data management is to identify and report the health condition of a patient before it gets complicated.
- Big data also helps us in identifying the patients, who accuses false claims by accessing large information.so it can be easily detected by reporting the fraud cases for evaluation.
- Because of its progressive innovations, it can easily detect and submit the false claims than any mankind.
- Physicians get certain patient data like some charts, particular facts regarding the subject by using traditional methods.
- By the use of big data, practitioner can easily know patients main cause of illness, thereby he can treat patients so that he can reduce some signs and symptoms.[8]

Disadvantages

- The main drawback of big data is one cannot have their privacy, as the medical documents which are kept to be secret by the individual is revealed by using big data.
- Privacy of a person is greatly affected by using big data, but it gives opportunity for physicians to treat their patients to the fullest.
- Due to the advanced technology, most of the people don't go to clinics for their individual check up and started treating their own treatment by following modern technology.[7]

Challenges of health analytics using Hadoop software

Cleansing

All the physicians and patients are interested in the cleanliness of the clinic and surgical suite, but they are not attentive in maintaining the data clean.

Data Cleansing – It is known as cleaning of data and it makes sure that the databases should not fluctuate, precise and appropriate information that should not violate.

Security

The first priority for all healthcare systems, especially in hacking of data, pulsation of prominent violations, blasphemer sequences. Data can be secured by HIPPA, which includes a list of specialized protection for organizations that stores PHI (Protected Health Information) which includes protocols, controlling the access, corroboration, monitoring, probity, conveyance safe.

Safeguards translates into common security approaches using present day Anti – virus software network, setting up of firewalls, encoding confidential information and several factors Attestation.

Health care systems should frequently remind their employees about essential data security protocols and should persistently review, who can access the valuable data resources in case of spiteful stakeholders.

Storage

As there is a rapid increase in the aggregate of healthcare information, the IT department faces challenges regarding critical price, surveillance and execution, but the front line clinicians don't have an idea of where their data is being kept.

The prices and influences of on ground data centres are no longer bearable by some suppliers. While many institution are on ease with on ground data storing as it assures surveillance check, entry and uptime.

While onsite server network is overpriced to calibrate, hard to support for and likely to cause information silos across different departments. Virtual storage has become an admired choice as it is affordable and trustworthy.

Governance

Health Care information has a long period of validity especially, in clinical field. Where patient's information can be available for minimum of 6 years, providers may need to utilize anonymous datasets for research projects, which makes existing governance and case management is the main concern.

Data can be reused, like quality measurement and performing comparisons. It is necessary for the researchers to know by whom and for what and who used the data previously.

To initiate complete, accurate, present day metadata is the core element for Data governance plan. It allows experts to recreate the former demands, which can carry out scientific research and for setting standards, it prevents from destruction of data.

To manage the improvement and mentorship of significant metadata. Data governance will make sure that elements have standard descriptions, configurations and documented auspiciously from establishment to elimination and stay beneficial for mission.

Querying

For any health care systems to query the information they need persistent databases and strong governance protocols to get expected answers. The capacity to inquire data is elementary for providing information and analytics.

Vigorous metadata and a good supervision of protocols make it comfortable for the institutions to query their information and acquire the awaited results. The skill of querying the information is a stepping stone for reporting and analytics. Before they can acquire a fruitful and meaningful analysis of the big information assets, the healthcare institutions should conquer the hardships regarding the data.

The first thing, they have to overcome is the information siloes and also the compatibility issues that obstructs the query tools from gaining entry to the institution's warehouse of information. If a dataset is stored in different formats or kept in a multi -walled off systems, then it is impossible to acquire a full profile of an institution's position and stature or health report of a sole patient.

Many institutions use structured Query Language (SQL) for big information sets and analogous databases, but this is useful only when a customer can rely on the information provided by them. So, an absolute out and out data with a great order should be provided

Visualization

A clinician can consume data from the cleanliness and engaging data visualization, so they can easily use it. Colour coding – most popular technique that provides an instant response (such as red-stop, yellow-caution, Green-go).

Examples – Bar graphs, pie- charts, histograms, bar diagrams.

Health care systems should review the good data presentations such as charts with proper sections and it should represent figures and their labelling information to reduce confusion.

Drawbacks – overlapping text, low- quality graphs, this irritate users to avoid data.

Updating

Health care systems are not fixed, they require frequent updates to present day, such as address/marital status- might change rarely in their lifetime.

It undergoes a great challenge for health care systems, to keep updating all the time, providers should have a clear idea of data about their updates and automation procedures without the damage of the quality information.

While adopting an update, clinicians should not confuse to access patient decision making.

Reporting

After the completion of questioning procedure, the contributor should produce a document which must be brief, convenient and understandable to the assembled spectators. Principles and reliability of data has a great effect on downbeat of the documented reports.

Lack of sufficient data may create a distrust report in the final step of the procedure. The provider should be able to interpret the contrasts between analysis and reporting.

In order to anal, always report must be required. Few reports turned out to be focusing on the latest popularity then summarize or gain the confidence of reviewer to take a particular action. Institutions must clearly know the idea of using their reports and make sure that database managers were able to produce required data.[8]

"HEALTH CARE SYSTEMS CAN BE RESTRUCTURED BY USING HADOOP SOFTWARE"

CONCLUSION

Health care systems, when executing a new idea into their framework, it should be very keen and profitable. Increased complications in the collection of large data in health sector gives more possibilities for Hadoop. Most of the hospitals are trying to figure out the best big data analytical tool in order to enhance better medical services for the patients and let them to take part in the prospective analysis and population health care administration. One of the drawback for this software is that it does not have complete equipped tools for data management, administration, descriptive data, data cleansing and standardization.

"A PERSONALIZED CARE BE PROVIDED BY THE CLINICIANS FOR THE PATIENTS BENEFIT"

REFERENCES

1. Kavitha. G, IV B.Tech IT, Dr. D.Prabha, Clinical Data Analytics in Big Data Using Hadoop, International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015, 24-27.
2. Alison Bolen, SAS Insights Editor, How do you know if you're ready for Hadoop?, available online at: sas.com/en_in/insights/articles/big-data/ready-for-hadoop.html
3. Elizabeth O'Dowd, CenturyLink, Cloudera Release Big Data-as-a-Service Solution, available online at: https://hitinfrastructure.com/news/centurylink-cloudera-release-big-data-as-a-service-solution

4. Tom White, Hadoop, the definitive guide, Yahoo Press, June 2009.
5. What is Clinical Data Management (CDM)?, available online at: https://www.mhaonline.com/faq/what-is-clinical-data-management
6. Healthcare applications of Hadoop and Big data, published on 16 Mar 2015, Latest Update made on May 01, 2016. Available online at: https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85.
7. Rick Delgado, The Use and Abuse of Big Data and Hadoop, published on February 26, 2014. Available online at: https://www.smartdatacollective.com/use-and-abuse-big-datahadoop/
8. Jennifer Bresnick, Health IT Integration Leaders Key to EHR, Big Data Success, published on March 02, 2016. Available online at: https://healthitanalytics.com/news/health-it-integration-leaders-key-to-ehr-big-data-success
9. Christine Taylor, Hadoop and Big Data: Still the Big Dog, by BICORNER.COM on July 27, 2015.

Source of Support: Nil, **Conflict of Interest:** None.