A Survey: Big Data Analytics in Health

Simardeep Kaur[1], Mainder Singh[2]
[1-2]*Department of Computer Science,*
Punjabi University, Patiala, India.

Abstract-*'Big-Data' is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of 'Big-data'. A major area of research of big data exists in healthcare field. The present study is an attempt to focus on classification of imbalanced huge amounts of dataset in medical field. However, issues related to accuracy, integrity exists in use of big data in medical field. Analytics of big data in healthcare can provide both clinical and management benefits. The present study will attempt to present Big-data, tools and approaches for analysis of health informatics to help advance understanding of medicine and medical practice.*

Keyword: *big data, health, mapreduce*

I. Introduction

Big Data is a term coined for the large and complex datasets or combination of datasets, such that traditional database and software techniques are inadequate to deal with them [1].Big data is a buzzword that describes enormous amount of structured, semi-structured and unstructured data that is complicated to perceive, acquire, manage and process by ConventionalIT, software/hardware tools and technologies, such as relational databases and desktop statistics within a tolerable time [1,5].

Tremendous amount of data that comes from variety of sources and in a variety of formats at an alarming velocity, is exploding day by day. This massive volume of data comes from everywhere: sensors data, social media data, public web, transaction records, machines log data, cell phones GPS signals to name a few. Around 280,000 tweets and more than 100 million emails are sent every minute. Google handles 2 million search queries and Facebook process 350 GB of data every minute [1, 9].The amount of data in all fields is growing exponentially. At least 2.5 quintillion bytes of data is produced everyday .Only in the last two years, 90% of the world's data has been created. Big data is everywhere and processing such huge amounts of complex and dynamic data can provide businesses real insights. Big data has the potential to benefit business with advanced analytics methods that extract value from the data, better operations and intelligent decisions [1].Big data is not just about finding insights from complex and voluminous data, also aims to answer the questions that were previously unanswered. The applications of big data include the areas such as social media, business, healthcare, science and research, banking, education, banking and so on [9].

A big-data revolution is under way in health care. Electronic health records (EHR), machine generated/sensor data, health information exchanges, patient registries, portals, and genetic databases, public records and so on are the major sources of big-data in healthcare area. Although complexities exist in healthcare data, still potential and benefit in developing and implementing big data solutions exist in this realm [2,3]. The big data's role in medical field is to build better health profiles and predictive models to diagnose and treat patients in a better way.Big data in healthcare is used to improve medical practices such as predict epidemics, cure disease, lowering costs, improve quality of life and avoid preventable deaths [4].

II. Big data in Healthcare

The volume of big data in healthcare field is predicted to increase over the coming years. The main reason behind the growing complexity and amount of data is due to the movement of the medical practice from ad-hoc and subjective decision making to evidence-based healthcare [6]. Effective evidence-based healthcare decisions needs collection, processing and interpretation of voluminous data. Big data in healthcare includes heterogeneous, multi-spectral, observations, Electronic health records (EHR), genetic databases and so on derived from different sources.However recent trend is towards the digitization of the healthcare data from stored printed form [10,7].

The Institute for Health and Technology Transformation (Newyork) estimates that in 2011, 150 Exabytes of data was produced by US Healthcare industry. The rapidly increasing data has lead to an expenditure of $1.2 trillion towards healthcare data solutions. Big data analytics in healthcare can help to reduce data management expenses by $300 - $500 billion [10].

Big data analytics is the process of collection, organization and analysis of large sets of data to discover patterns and provide useful insights. The sheer volume and format of healthcare data is a big analytical challenge [8]. Big data analytics in healthcare is to make use of tools and techniques that can leverage voluminous data effectively [10]. Big data challenge in healthcare is to infer knowledge from complex heterogeneous patient sources, effectively handling large volumes of medical imaging data and capturing patient's behavioral data through several sensors. Major challenge is to provide personalized care to the patient, provide right intervention to the right patient at right time [6].Data analytics in healthcare can be used to raise the standards of public health, Electronic Medical Record (EMR), patient profile analytics, Genomic analytics, Fraud analysis, safety monitoring. Data analytics in healthcare is equally effective ranging from individual physician to multi provider healthcare organizations [9].

III. 4 Vs of Big Data:

Big data analytics in healthcare is associated with four primary characteristics or dimensions of big data: volume, velocity, variety and veracity.

A. ***Volume 1st V:***Volume of big implies massive amount of data. Although there is no fixed size of data that can be considered as a threshold to be termed as big data, however term refers to massive volume of data difficult to handle with traditional tools and techniques. In 2012, EHR generated 500 petabytes of data that is expected to grow up to 25000 petabytes [11, 12]. Today, much of the enthusiasm is to find useful insights from new and more extensive sources of data. Major sources of big data in healthcare industry are: clinical records,Electronic health records (EHR), machine generated/sensor data, health information exchanges, patient registries, portals, and genetic databases, public recordsand health research records [2,3]. Data in healthcare industry at such a massive scale has the potential to transform healthcare. For instance, a 3D CAT scan takes up 1 GB and human genome takes up 3 GB of data [14].

B. ***Velocity 2nd V:***velocity is the measure of the speed at which data flows in from sources and is considered as a hallmark of big data. The speed of the data generated by patient encounters and patient monitors is increasing.Big data in healthcare is available real-time and often arriving in bursts rather than at a constant rate [14].

C. ***Variety 3rd V:*** Variety refers to the diversified sources and types of data. Big data comes in variety of forms and formats such as structured, semi-structured and unstructured. Structured data forms only 5 to 10% of the data. Structured data is the simplest way to manage data in rows and columns in databases.Semi-structured data is also 5 to 10% of data. Unstructured data is around 80% of the data. Unstructured data includes e-mails, photos, videos, audios, webpages and many other documents. Digital capture and management of diagnostic imaging requires special data formats [13].

D. ***Veracity 4th V:***Veracity of data refers to the noise, abnormality and error-freeness of data. Different data sources vary in credibility and reliability of data. Unstructured data that forms most part of the data is highly variable and is often incorrect. Healthcare analytics should aim to find useful insights from such data to cure patientsand to make better decisions [13].

IV. Big Data Analysis Tools:

Nowadays, focus is to understand the meaning and importance of the data rather than just the collection of data. Data analytics refers to process of application of algorithms to analyze data sets and extract useful patterns and information.The very first requirement of big data analytical processing is to load the data at fast speed i.e reduce the load time. Second requirement for big data analytics is fast query processing. Third requirement is the efficient utilization and management of the storage space. The fourth requirement of big data analytics is to adapt highly dynamic workload patterns [15].

Map Reduce is basically a parallel programming model, which is suitable for big data processing. Map Reduce paradigm refers to scaling out rather than scaling up. Map Reduce breaks down the task into stages and then these stages are executed in parallel to reduce overall completion time of the task [15].

Big data sizes are increasing constantly ranging from Terabytes to many Petabytes. Such voluminous amount of data is difficult to capture, store, search, share, analysis and visualize. The larger the set of data, the more difficult it becomes to manage and a need arises for new tools and methods for analytics. Most of the tools of the big data are based on Hadoop architecture which provides reliability, scalability, and manageability

by using map-reduce paradigm [16]. Several tools for processing big data are discussed below:

- ✓ **Apache Hadoop:** Hadoop is a software framework that permits large scale distributed data analysis using MapReduce programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Two main components of Hadoop are:
  - A storage part known as HDFS and
  - A processing part known as MapReduce.
  - **HDFS:**The Hadoop Distributed File System (HDFS) is a distributed file system designed for applications that have large data sets. HDFS has a master/slave architecture. HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data [16].

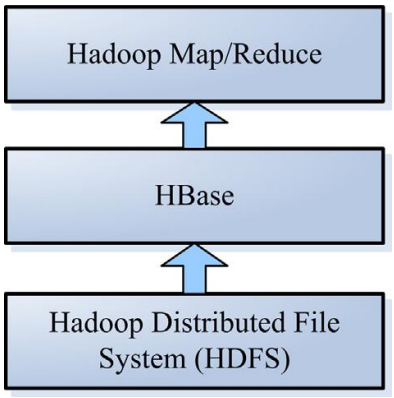

Figure 1: Hadoop system architecture

- ✓ **NameNode:** The NameNodeis centerpiece of HDFS also known as Master, executes HDFS operations like opening, closing and renaming files and directories. It also helps in determining mapping of blocks to DataNodes. NameNode manages the metadata for the cluster such as list of HDFS files, state of the file and the access control information [16].
- ✓ **DataNode:** The DataNode also known as Slave, is responsible for storing the actual data, serving read and write requests from HDFS file system's clients. The client sends data directly to and reads directly from DataNodes and never flows through the NameNode [16].
- ✓ **Map-Reduce:** It is the essential component of Hadoop software framework. Two important tasks of mapreduce are: Map and Reduce. Map takes the data set and broke it down into tuples. Reduce part performs summary operation and is always performed after Map part. It considers Map part's output as input and combine data tuples into smaller set of tuples [17].

V. Conclusion:

In this study, we have provided the brief overview of big data, big data in healthcare, big data analytics in healthcare, 4 V's of big data in the light of healthcare informatics, architecture of big data analytics and about Hadoop tool to process large data sets of healthcare industry. Big data provides the ability to track trends and patterns from multiple sources of health data. Big data solutions in healthcare attempts to cost-effectively solve the challenges of large and fast growing data volumes.Big data analytics in healthcare would be fruitful in terms of Genome processing and DNA sequencing, personalized treatment planning, assist diagnosis and monitor patients.Big data has many implications for patients, providers, researchers, payers and other healthcare constituents.

REFERENCES

[1] VibhavariChavan and Rajesh N. Phursule, "Survey Paper on Big Data", International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014, 7932-7939.

[2] Ashwin Belle, RaghuramThiagarajan, S.M. Reza Soroushmehr, FatehmehNavidi, Daniel A. Beard and KayvanNajarian, " Big Data Analytics in Healthcare", Hindawi Publishing Corporation BioMed Research International Volume 2015, Available from:http://dx.doi.org/10.1155/2015/370194.

[3] Available from: https://www.verywell.com/sources-of-big-data-in-health-care-1739184.

[4] Available from: https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#66177bb02873

[5] Harshawardhan S. Bhosale and Devendra P. Gadekar, "A Review Paper on Big Data And Hadoop", International Journal of scientific and research publications, Volume 4, Issue 10, October 2014.

[6] Jimeng sun and Chandan K. Reddy, "Big Data Analytics For Healthcare", SIAM international Conference on Data Mining, 2013, Available from:http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare.

[7] Ivo D. Dinov, "Volume and Value of Big Data Healthcare Data", Published in J Med

[8] Stat Inform, 2016.

[9] Available from: http://webopedia.com/TERM/B/big data analytics.html

[10]Available from: http://www.builtinla.com/blog/significant-benefits-big-data-analytics-healthcare-industry

[11] Available from: https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85

[12] Feldman, B., Martin, E.M. and Skotnes, T, "Big Data in Healthcare Hype and Hope" Dr. Bonnie 360, 2013, pp. 19.
[13]Feldman B, Martin E, Skotnes T.," Big data in healthcare hype and hope" GHDonline, 2012.
[14]Ahmed Abbasi and Roger H.L. Chiang, "Big data research in information systems:Toward an inclusive research agenda", 2016.
[15]AtreyiKankanhalli, jungpil Hahn, Sharon Tan and Gordon Gao,"Big data analytics in healthcare: introduction to special section", InfSyst Front, 2016.
[16]Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", 2016.
[17]Available from: http://itm-vm.shidler.hawaii.edu/HDFS/ArchDocOverview.html
[18] Jens Dittrich and Jorge-Arnulfo Quiane-Ruiz, "Efficient Big Data Processing in Hadoop MapReduce", 2015.