INTERNATIONAL JOURNAL OF STATISTICS AND MEDICAL INFORMATICS

Vol. 9 No. 1/2018, pp: 1-6

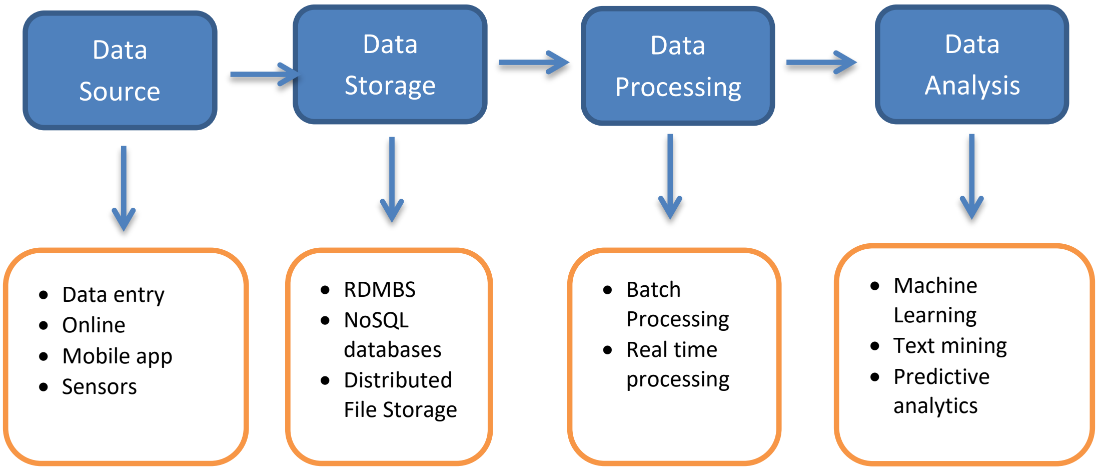
Big Data and its application in Biomedical Domain

Editor, International Journal of Statistics Medical Informatics (IJSMI)
editor@ijsmi.com

1. Introduction

Big data [1,2,3] redefined the analytical space with its features such as Volume, Velocity, Variety, Veracity and Value. Due to advancement in the technology, storing, retrieving of data is becoming easier than in the previous decades. Due to this fact, the volume of data in the form of text, image, sound is increasing at a rapid pace in all the fields including the biomedical field [4, 5]. Patient records in the form of electronic medical records are being stored in the data cloud which includes data in different forms. To analyse the data of such magnitude and variety, traditional analytical tools and methods are not sufficient. The Big Data analytical techniques help to analyse these kinds of data and helps us to arrive at decisions quickly. Managing the privacy, security and government regulations related to patient data remains as a challenge [6] in implementing Big Data analytical tools in biomedical domain. This paper starts with the overview of Big Data architecture and moves on to explaining the tools and technologies used in Big Data and the uses of Big Data in Biomedical Field.

2. Big Data Architecture



The traditional data analytics architecture require the data to be moved to computing environment for data analysis purpose which require more time, repetition of data and create congestion in the network.

Big Data architecture [1,3,7] helps us to overcome the above problems through parallel processing at the distributed environment. Big data architecture involves data source, data storage, data processing and data analysis.

a. Data sources can be through manual data entry, online, mobile apps, logs files, sensors in the form text, sound, image, videos etc.
b. The collected data can be stored and cleaned in the Relational Data Base Management System (RDBMS), Not only SQL databases (NoSQL) and Distributed File Storage (DFS).
c. The stored data can be processed by using batch, real time and hybrid processing.
d. The data can be analysed using machine learning, deep learning, text mining and data mining techniques.
e. Big data architecture includes mechanism for monitoring, ensuring security and providing Quality of Service.

3. Big Data tools and technologies

Some of the important tools and technologies used in the Big Data architecture are given below:

a. **Apache Hadoop**[8]

   It contains two components one is Hadoop distributed File System (HDFS) and another is MapReduce [9].

   Hadoop distributed File System helps to store the file across the distributed systems in the cluster and enables to access the files easily from the systems. The nodes (system) operate parallel in the clusters.

   For example patient records can be stored individually across the hospital systems which are connected as nodes in the Hadoop File Processing system. The same patient has records stored in different hospitals servers which can contain different content related to that particular patient across the cluster.

b. **MapReduce** enables distributed processing of data across different systems (servers) in the cluster. It involves two tasks first task is Map and then the second task is Reduce. The data in the system are converted into tuples by Map having key and value pairs and Reduce combines the tuples into smaller number of tuples and produces the final result.

For example the patient data is stored with two columns Patient ID, Haemoglobin at different servers across the Hadoop System. Each server contains multiple values of Haemoglobin for a patient and we would like to compute the max value of Haemoglobin for each patient whose data is stored across the systems in the cluster. The Map function computes the max value of Haemoglobin for each patient from each server and stores as tuples (key – Patient id, value – Haemoglobin value). Here each patient might have two or more tuples (one or more max value of Haemoglobin) associated with a patient as the patient record might be stored in more than one hospital. The Reduce function combines or aggregates the tuples calculate the max value of Haemoglobin for each patient. The final result will contain only one max value for each patient.

c. Apache Hive [10]

Apache Hive is a data ware house software that facilities performing of different operations (reading and writing) and management of large datasets in the distributed system environment

d. Apache Spark [11]

Apache Spark is a unified analytics tool for large scale data processing. It can handle data sources from sources and platforms such as Java, Python, R and SQL. It includes libraries for SQL data structures, machine learning and graphical and stream processing.

e. **Apache Storm** [12]
   Apache Storm is a real time data processing technology which works with different programming languages and is faster than Apache Hadoop Batch processing technology.

f. **Apache HBase** [13]
   Apache HBase is the Hadoop distributed and scalable non-relational database suitable for sparse datasets.

g. **NoSQL** databases [14]
   NoSQL databases help to store non-relational data and unstructured data which are stored as pairs (key and value) and can be retrieved without using Standard Query Language (SQL).

h. **Machine Learning [15] and Text Mining Techniques [16]**
   Machine Learning techniques involve analysing data using supervised and unsupervised techniques to carry out predictive and classification analysis. Text mining involves mining of text data using text mining models and Natural Language Processing models

4. Uses of Big Data in biomedical domain

The volume of data in the biomedical domain[17,18,19,20] is increasing in the faster rate in the forms of electronic medical records, radiological images, data related to clinical trials and research, genomic sequence, social network, data generated from biomedical sensors, log files and mobile apps. These data are generated in different forms such as numeric, text, image, sound and video. Big Data analytical techniques help to analyse such volume and variety of data but at a higher speed and accuracy.

The following are the uses of Big Data analytics

a. Analysing the disease outbreaks real time
   Epidemiologist are able to detect the disease outbreaks from the real time data generated from different sources such as websites, social media, search engines and maps.

b. Big Data analytical tools help the clinicians to provide personalized clinical care through mobile applications and biomedical sensors.

c. Linking different sources of data in the clinical domain such as clinical registries, clinical trials and clinical research findings and electronic medical records will help the clinical decision makers to take a holistic decision.

d. Big Data analysis can contribute to Evidence Based Medicine an important field of clinical research by including both structured and unstructured data analytical capabilities.
e. Detection of frauds in health insurance claims
f. Big Data analytics can speed up the process of drug discovery and clinical trials conducted by the clinical research and Pharmaceutical organizations.
g. Predictive and pattern discovery analysis in clinical domain. Predictive analytical in terms of predicting the survival rate, adverse reaction, Length of Stay,
h. Unfolding the large volume of genomic sequence

5. Conclusion

The paper provided an overview of Big Data, its architecture, tools and technologies used in Big Data and the uses of Big Data in Biomedical Domain. The Big Data Analytics has changed the way the data is being stored, processed and analysed but the challenges related to security and privacy of patient data remains to be addressed.

6. Reference

1. Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.
2. Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing, 74(7), 2561-2573.
3. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.
4. Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3-13.

5. Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352.
6. Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and challenges of big data computing in health sciences. Big Data Research, 2(1), 2-11.
7. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2, 652-687.
8. https://hadoop.apache.org/
9. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
10. https://hive.apache.org/
11. https://spark.apache.org/
12. http://storm.apache.org/
13. https://hbase.apache.org/
14. Pokorny, J. (2013). NoSQL databases: a step to database scalability in web environment. International Journal of Web Information Systems, 9(1), 69-82.
15. Editor, IJSMI (2018), Machine Learning: An overview with the help of R software, ISBN 978-1729293577
16. Editor, IJSMI (2018). Application of statistical tools in biomedical domain: An overview with help of software. ISBN 9781986988551.
17. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 33(7), 1123-1131.
18. Costa, F. F. (2014). Big data in biomedicine. Drug discovery today, 19(4), 433-440.
19. Sun, J., & Reddy, C. K. (2013, August). Big data analytics for healthcare. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1525-1525). ACM.
20. Kuo, M. H., Sahama, T., Kushniruk, A. W., Borycki, E. M., & Grunwell, D. K. (2014). Health big data analytics: current perspectives, challenges and potential solutions. International Journal of Big Data Intelligence, 1(1-2), 114-126.