TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

Big Data Analytics System for Health Care: A Literature Review

Vaibhavi S. Ghelani*, Prof. Purvi Ramanuj

*[1] M.E. Information Technology, L. D. College of Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India
vaibhavighelani@gmail.com

[2] Assistant Professor, Information Technology, L. D. College of Engineering, Gujarat Technological University, Ahmedabad, Gujarat, India
purviramanuj@gmail.com

Article Info
Volume 83
Page Number: 542 - 548
Publication Issue:
July - August 2020

Article History
Article Received*: 06 June 20*
Revised*: 29 June 2020*
Accepted*: 14 July 2020*
Publication*: 25 July 2020*

Abstract

Health is generating enormous volumes of data that can provide vital understandings into medical and real features of healthcare delivery. There is a general dearth of specific and combined health data analytics platforms that offer technical approaches to support the entire health data analysis channel. For example, health data selection, integration, inspection, visualization, and distribution. This paper reviews the numerous practical architectures of a health data analytics platform that offers a technical solution for analyzing big' health data originating from several sources with varied terminologies and plans. The present study which stresses the use of the enormous size of health information while merging multidimensional material from varied sources is discussed. Finally, a Research Gap is found in all studied technical architecture and probable solutions proposed.

Keywords: Big Data, Healthcare, Health informatics, Big health data, Hadoop, Data integration, Data standardization

I. INTRODUCTION

Big Data is a term for data set that are very much big or intricate that conventional data processing applications are inadequate to handle them [1]. In today's world, the rate of data generation has increased exponentially by increasing the use of data-intensive technologies. To handle this data, we require different approaches: Different Methods, Tools and Architecture which can be referred to as Big Data Analytics. The key organizations to implement Big Data are eCommerce and newly started businesses. Companies like Google, eBay, Linkedln, and Facebook are using Big Data Analytics from their commencement.

II. BIG DATA APPLICATIONS

A. Government Sector

The governmental sector uses Big data to rise competences in terms of value, output, and invention. In government following are the major departments which use Big Data Analytics:

- Cyber Security and Intelligence
- Crime Prediction and Prevention
- Pharmaceutical Drug Evolution
- Weather Forecasting
- Tax Compliance
- Traffic Optimization

B. Health Care

Big data analytics have improved healthcare by providing person-specific medical treatment according to a person's medical history. Age, Demographics, etc., Researchers are gathering the data to see what actions are more functioning for specific circumstances, recognize patterns related to medication side effects, and gains other significant data that can help patients and reduce costs.

C. Manufacturing

The major benefits of Big Data applications in the manufacturing industry are:

- Maintenance of Product quality
- Supply planning
- Production process defect measurement

Published by: The Mattingley Publishing Co., Inc.

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

- Output forecasting
- Increasing energy
- Simulation of processes
- Provision for bulk-customization of production

D. Media and Entertainment

The Major benefits in the Media and Entertainment industry are:

- Forecasting what the audience chooses
- Scheduling optimization
- Manufacturing process defect tracking
- Growing customer attainment and retention
- Ad Targeting
- Content Monetization and New Product development

E. Internet of Things

Information gathered from ***IoT*** devices provides scheduling of device interconnectivity. Such mappings have been used by diverse companies and governments to raise competence. IoT is also used as a medium for gathering physical data, and this physical data is used in health and industrial contexts.

F. Education

In the education industry the main uses of Big Data are:

- Improve student result
- Customize Programs for individual student
- Reduce Dropouts
- Targeted International recruiting

G. Transportation [36]

- Many train operating organizations that have already started using big data to process the obtainability of seat data in real-time and also to notify the passengers waiting on platforms about the carriages having the greatest number of available seats.

- In transportation, the effective study of repeat grievances made by a single customer numerous times through big data could result in a more effective response. Thus, it helps in offering advanced solutions like smartphone technology to solve a variety of issues.

- Big data can be used to eliminate errors and decrease unnecessary spending. It can be used to find the problems associated with delays and downtimes for transport upkeep.

III.FUNCTIONAL ARCHITECTURE

The functional architecture of Big Data model can be divided into four different stages:

A. Health Data Finding [28] and Assortment

This functionality allows users to interactively view /search and hand-picked the relevant data.
Sources for this include following:

1) Social media sites like Facebook, Twitter, etc: People are turning to social media for health-related matters like getting information regarding health care products and services, sharing experiences and seeking expert opinions. A lot of health-related messages are often communicated through these sites and several research works are done in the analysis of social networks for effective health care [31].

2) Medical and health insurance information related websites.

3) Machine Specific data (Machine to Machine, Machine to Human, Human to Machine): Readings from remote sensing devices, meters, satellites, IoT devices. Medical small sensors can be worn on by the human body. These sensors measured the data and sent it to the external medical doctor. With these sensors, the patient can move from one place to another. Many people die from different fatal diseases when it is diagnosed lately. For this purpose, the Wireless Body Area Network is used to detect disease early and prevent fatal diseases [37-40].

4) Health Operational data: Medical reports billing information, insurance files.

5) Biometric data as X-ray, fingerprint, palm prints, genetics, retinal scans, handwriting, including themedical images also.

Once the data has been collected it has converted into a type that is suitable for further processing. A service- oriented architectural approach together with web services is one example of transforming the data [12].

We can also use the XML Configuration Data Store for storing source datasets for data integration. The XML configuration data store also contains integration metadata, connection, and user credential information for all available external source datasets [30].

Published by: The Mattingley Publishing Co., Inc.

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

In data warehousing, data is taken from various sources and is made ready for processing. Through the various steps of extracting, transforming, and loading (ETL) the data from diverse sources is cleansed and readied [28].

Fig. 1 Functional Architecture of Big Data Model

B. Health Data Understanding and Addition:

Health data is collected by heterogeneous communities for specific programs; therefore, it is important to understand the etiology of the data—i.e. who created, through which processes and for what purpose. These functionalities take into account the purpose and provenance of health data to

1) Standardize the schema and content of local databases which is also called semantic interoperability. Semantic interoperability implies that the data originating from multiple sources upon integration maintains its correct meaning and is standardized to a medical terminology system.

2) Integrate the standardized local datasets to generate a question-specific focused dataset. Integration is achieved at the schema level where either complete tables or select columns can be integrated. The etiology of the data is used to determine the intent of the data to ensure datasets/attributes with similar intent is integrated. Standard data privacy protocols are employed, no patient identifiers are present in the data [28].

C. Health Data Analytics

In this section, we can make several decisions for analysis of the data input technique, distributed design, platform selection, and models.

Different fields such as computer science, statistics, economics, and applied mathematics have developed avariety of techniques and tools for analysis of information [43].

Hadoop is the best tool for the collection and aggregation of web search indices. Hadoop is an open-source of distributed data processing. This platform belongs to the class "NoSQL" technologies while others include CouchDB and MongoDB and many more that were advanced to process the big data in distinctive ways. Hadoop handles extremely large data set by allocating the task to numerous nodes. Each node solves different parts of the large database and program and then collects the outcomes together.

Hadoop [3] is being used in an enterprise as a data controller and processing tool. Either the data is structured or unstructured. The adjacent ecosystem of extra platforms and tools supports the Hadoop distributed platform [5].

We can do advance analysis in Hadoop with the use of Impala and RImpala for the execution of distributed queries [30]. Impala a massively parallel processing (MPP) SQL query engine that runs natively in Apache Hadoop [35], selects data stored on HBase and HDFS. RImpala is utilized to establish the connection between R and Impala through JDBC integration.

Data mining methods can be used to discover new and interesting patterns, inter-attribute correlations and associative and causal relationships [29].

We can use statistical methods to present summaries, aggregations, charts, and reports. The open-source R statistical environment is exploited to deliver informational analytics capabilities [29].

D. Health Data Visualization

For health decision-makers, the ability to understand, interact and respond to data analytics results is paramount. This functionality offers a range of data visualizations— such as geospatial maps, dashboards, scorecards, 2D and 3D models and animations—to present the analytics results. The four important areas of big data analytics in the healthcare system contain queries, reports, OLAP, and data mining. Visualization embraces all the above- mentioned applications.

In particular, the visualization targets the needs of decision-makers who need to drill down/up for further insights.

TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

TABLE I
Comparison of Major Reviewed Papers

Attribute	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5
Focus	The huge volume of medical data: combining Multidimensional data from different sources.	To design analytical experiments: perform complex analytics on health data.	Data standard-ization from multiple source using SNOMED CT	A real-time remote health status prediction system using machine learning model	A healthcare system based on energy harvesting technique welcomes real-time and offline data
Big Data Transformation Technique	A service oriented architectural approach together with web services, data warehousing [44]	Semantic inter-operability methods [7-8] in conjunction with medical terminologies of SNOMED CT, LOINC and ICD-10 [9]	1. Data Standard-ization Component: Semantic inter-operability among data [8] 2. Data Integration Component	Incoming data streams are divided into small batches: processed by the Spark engine.	Pre-processing Layer: Redundant data along with the erroneous data is removed using Doppler and Specan methods [21,22]
Big Data Processing and analysis	Hadoop, CouchDB. MongoDB [3,5]	Exploratory Analytics, Semantic Analytics, Predictive Analytics, Information Analytics	Data Selection Planner Component, Data Analysis Component	Discretized streams-the sequence of RDDs, represents a continuous data stream. Operations on DStreams: converted to basic RDD transformatio n-computed by the Spark engine	Hadoop Processing Layer: Load Balance, Raw Data Storage, Message Queue, Hadoop Ecosystem, Parallel Processing
Visualization & Decision Making	Queries, reports, OLAP, and data mining	Geo-spatial maps, dashboards, scorecards, 2D-3D models and animations	Data Visualization Component: Data Selector selects data stored in HBase linked with Hive store and invokes the Data Selection Executor	Spark's machine learning, MLlib, supports decision tree algorithm, used to build a scalable machine learning model	Data Application Layer: Accessor results in the storage device, decision- making unit, the communication medium

Published by: The Mattingley Publishing Co., Inc.

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

IV. CONCLUSION

The central idea of this review paper is to analyze the most necessary part of the functional architecture of the Big Data analytics system in health-care. Table I conclude important features that are implemented in five main paper reviewed. It gives a brief idea of different methodology which can be captured in various parts of the Big Data model. It saves a considerable amount of time and money in using existing technologies to its full potential than developing new ones to do the same job. Through further enhancement of already developed framework, we can predict the presence of a variety of diseases.

Experts in this field need to understand the importance of Big Data Analytics in health-care. Experts need to analyze requirement specific Big Data Model instead of using a generic model.

It is important for software expert to understand than health-care data is different from other data and need to be taken care and analyzed with utmost accuracy.

As future work, this application can be linked with the systems of healthcare providers to form a complete real-time health care system. An energy harvesting technique should be a must that prolongs devices lifetime.

V. FUTURE WORK

In existing technical architecture, we can append a strong Decision-making unit that is supported by the intelligent medical system, machine learning and other medical problems for analysis and making a decision. We can provide greater insights to healthcare providers to improve patient care, cost optimization and better outcomes from the history of human beings, a current prevailing disease in his/ her demographics, clinical guidelines, diagnostic support.

For example, we can use a decision support system for the following:

1) Decision Support System can implement rules and patterns for individual patients, based on clinical parameters, and raise warning flags when such rules are violated. We can save lives with clinical interventions as and when these flags are generated. For example, for chronically ill patients, a deviation noticed by Decision Support System in, say, a heart-beat rate reading from a heart patient could result in an intervention before the patient gets into difficulty.

2) Decision Support System can be used to research patient populations to calculate the risks of non-compliance to prescribed management plans for individual patients.

3) Decision Support System can be used to generate knowledge bases to improve system-wide efficiencies and patient outcomes in learning healthcare systems.

4) Decision Support System can perform regular clinical decision prescribed medications and/or dosages for the patient.

5) Decision Support System can spot incompatibilities between prescribed medications and/or dosages for the patient.

6) We can use various machine learning Here is a glimpse of model design for decision support system using machine learning algorithm classifiers for making a decision.

A. **Model Design using a machine learning algorithm for decision support system:**

Fig 2. Shows how we can develop a model for decision support systems using a machine learning algorithm. We need to take medical data from a variety of sources, expert knowledge base and clinical parameters as inputs which may be called our training data. We can apply a machine-learning algorithm to the training data set and develop the classifier. Then we can take new medical data, test them using this classifier.

We can check the result of the classifier with various parameters (Example: Entropy, Gini Index, R^2, Adjusted R^2). If our confidence level is high for the given result, we pass on the related information, alerts, and warnings to all stakeholders (patients, doctors, etc.). If our confidence level is low, we again check our data with different classifiers developed by a different machine learning algorithm.

Take an example of disease – diabetes in females. We can take medical data from different age groups. Also, we need to take medical data of the same person before and

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

after lunch/supper. As per expert knowledge, the major

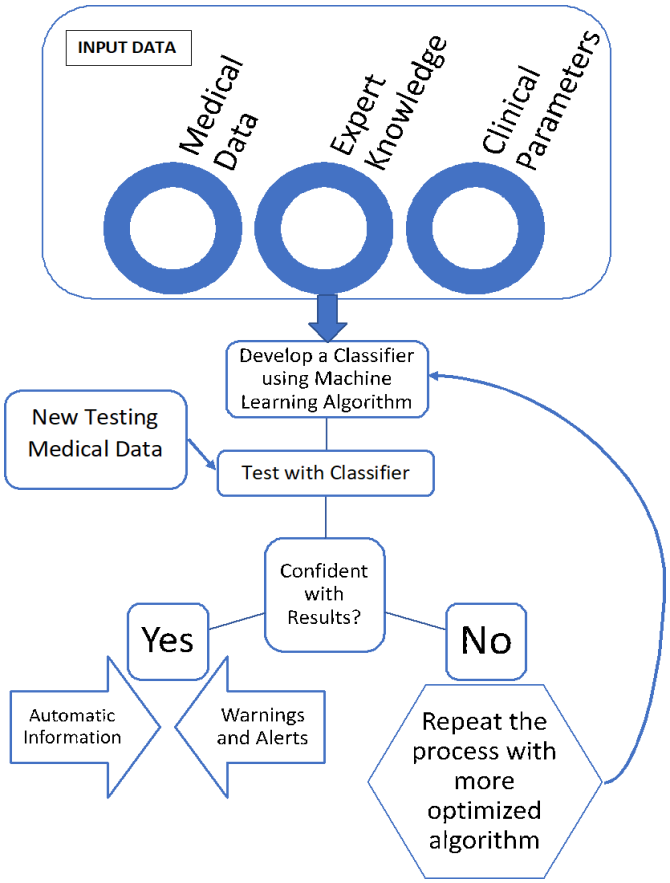

Fig 2. Model Design for Decision Support System

factors that affect any female's blood glucose are age, BMI, Pregnancies, Family History, BP, Glucose, Triceps, Insulin.

Now as per the clinical parameters' persons having glucose level >=126 before lunch/supper and persons having glucose level >=200 after lunch/supper can be considered as diabetic.

For example, we have taken Pima Indians Diabetes data from Kaggle.com which contains all our required feature data (Age, BMI, Pregnancies…).

We can apply any machine learning algorithm to determine which features can act as a classifier. The widely used machine learning algorithm in such a scenario are K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Gaussian Naïve Bayes and Random Forest.

Suppose we use K-Nearest Neighbors, then we split the data set in K equal parts. Here we take one-fold as training data and other folds as testing data. Then the other fold as testing data. We will follow this test in K-times. We will take the average accuracy of each process and consider it as final accuracy. Suppose we have received 70% accuracy through this model.

If we are not satisfied with this accuracy, we will again develop a new classifier with another machine algorithm like logistic regression. It may happen accuracy of this classifier may be above 75%. If we satisfy with this accuracy, we will test our new medical data with this classifier. We will give important information, alert and warnings to patients and/or their doctors who are classified as probable diabetic

REFERENCES

1. https://www.sas.com/en_us/insights/big-data/what-is-big data.html.
2. lliTT: Transforming Health Care through Big Data strategies for leveraging big data in the health care industry.
3. Borkar YR, Carey MJ, Chen L: Big data platforms: what's next? ACM Crossroads 2012,19(1):44-49.
4. https://www.marketquotient.com/blog/big-data-analytics-the-next-big thing/.
5. Zikopoulos PC, DeRoos D, Parasuraman K, Deutsch T, Corrigan D, Giles Harness the Power of Big Data. McGraw-Hill: The IBM Big Data Platform; 2013.
6. Ohlhorst F: Big Data Analytics: Turning Big Data into Big Money. USA: John Wiley & Sons; 2012.
7. WA. Khan, AM. Khattak, M. Hussain, MB. Amin, M. Afzal, C. Nugent, S. Lee, An adaptive semantic-based mediation system for data interoperability among health information systems. Journal of Medical Systems, Vol. 38(28), 2014.
   A. Ryan, P. Eklund, "A framework for semantic interoperability in healthcare: A service-oriented architecture based on health informatics standards," in Studies in health technology and informatics, vol. 136, eHealth Beyond the Horizon, 2008, pp. 759-764.
8. IHTSDO, History: SNOMED, www.ihtsdo.org/aboutus/ history/snomed/ , 2007.frontiers,'' Science, vol. 355, no. 6324, p. 489, 2017.
9. Apache Mahout, https://mahout.apache.org/
10. The R project for statistical computing, http://www.r-project.org/.
11. R. J. Kate. Towards converting clinical phrases into SNOMED CT expressions. Biomedical Informatics Insights 6(Suppl 1), pp. 29. 2013.
12. Shiny by RStudio, http://shiny.rstudio.com/.
13. The R project for statistical computing, http://www.r-project.org/.
14. Dredze M. How social media will change public health. Intell Syst IEEE 2012;27(4):81–4.
15. Denecke K, Krieck M, Otrusina L, SmrzP DologP, NejdlW, et al. How to exploit twitter for public

July – August 2020
ISSN: 0193-4120 Page No. 542 - 548

health monitoring. Methods Inf Med 2013;52(4):326–39.
16. Lee K, Ankit A, Alok C. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining; 2013.
17. Trigo JD, Eguzkiza A, Martínez-Espronceda M, Serrano L. A cardiovascular patient follow-up system using Twitter and HL7. In: Proceedings of computing in cardiology conference (CinC); 2013.
18. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. HotCloud 2010;10(10-10):95 http:// www.bbc.com/news/uk-20851797
19. Boyi Xu, Li Da Xu, Senior Member, IEEE, Hongming Cai, Cheng Xie, Jingyuan Hu, Fenglin Bu, Ubiquitous data accessing method in IoT-based information system for emergency medical services, IEEE Trans. Ind. Inf.10 (2) (2014) 1578–1586, http://dx.doi.org/10.1109/TII.2014.2306382.
    A. Paul, A. Ahmad, M.M. Rathore, S. Jabbar, Smartbuddy: defining human behaviors using big data analytics in social internet of things. IEEE Wireless Communications 23 (5), 68-74.
20. Ahmad Awais, Anand Paul, Mazhar Rathore, Hangbae Chang, An efficient multi-dimensional big data fusion approach in machine-to-machine communication, ACM Trans. Embedded Comput. Syst. (TECS) 15 (2) (2016) 39.
21. Attila Reiss, Didier Stricker, Introducing a new benchmarked dataset for activity monitoring, in 2012 16th International Symposium on Wearable Computers (ISWC), IEEE, 2012, pp. 108
22. Attila Reiss, Didier Stricker, Creating and benchmarking a new dataset for physical activity monitoring, in Proceedings of the 5th International Conference on PErvasive Technologies Related To Assistive Environments, ACM, 2012, p. 40.
23. Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, Claudia Villalonga, mHealthDroid: A novel framework for agile development of mobile health applications, in Ambient Assisted Living and Daily Activities, Springer International Publishing, 2014, pp. 91–98.
24. M. Mazhar Rathore, Awais Ahmad, Anand Paul, Jiafu Wan, Daqiang Zhang, Real-time medical emergency response system: Exploiting IoT and big data for public health, J. Med. Syst. 40 (12) (2016) 1–10.
25. https://www.edureka.co/blog/big-data-applications-revolutionizing- various-domains/
26. Big Data Analytics for Health Systems - 2015 *International Conference on Green Computing and Internet of Things* (ICGCloT) ©2015 IEEE.
27. H-DRIVE: A Big Health Data Analytics Platform for Evidence-Informed Decision Making - 2015 IEEE International Congress on Big Data.
28. Towards a 'Big' Health Data Analytics Platform - 2015 IEEE First International Conference on Big Data Computing Service and Applications.
29. L.R. Nair et al., Applying a Spark-based machine learning model on streaming big data for health status prediction, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j. compeleceng.2017.03.009 © 2017 Elsevier Ltd.
30. S. Din and A. Paul, Erratum to "Smart health monitoring and management system: Toward autonomous wearable sensing for the Internet of Things using big data analytics" [Future Gener. Comput. Syst. 91 (2019) 611–619]'', Future Generation Computer Systems (2019), https://doi.org/10.1016/j.future. 2019.06.035. © 2019 Elsevier. RImpala package, http://www.inside-r.org/node/223456.
31. Apache HBase, http://hbase.apache.org/
32. Doukas Charalampos, Ilias Maglogiannis, Vassiliki Koufi, Flora Malamateniou, George Vassilacopoulos, Enabling data protection through PKI encryption in IoT m-Health devices, in Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on, IEEE, 2012, pp. 25–29
33. https://www.businesswire.com/news/home/20180914 005108/en/Top- Benefits-Big-Data-Transportation-Industry.
34. S.M.R. Islam, D. Kwak, M.H. Kabir, M. Hossain, K.-S. Kwak, The internet of things for health care: a comprehensive survey, IEEE Access 3 (2015) 678–708.
35. Ghamari Mohammad, et al., A survey on wireless body area networks for e-healthcare systems in residential environments, MDPI Sens. (2016).
36. Cavallari Riccardo, et al., Survey on wireless body area networks: Technologies and design challenges, IEEE Commun. Surv. Tutor. 16 (3) (2014).
37. Sharma Raju, et al., Wireless body area network – a review, Int. J. Eng. Sci. (2016).
38. RImpala: R and Impala, http://cran.rproject.org/web/packages/RImpala/index.html
39. https://www.kaggle.com/saurabh00007/diabetes.csv.
40. Zhang, Jianguo. "Big data issues in medical imaging informatics", Medical hanging 2015 PACS and hanging Informatics Next Generation and Innovations, 2015.
41. Raghupathi W, Kesh S: Interoperable electronic health records design: towards a service-oriented architecture. eService Journal 2007, 5:39-57.

Published by: The Mattingley Publishing Co., Inc.