



IMPLICATIONS OF BIG DATA ANALYTICS IN DEVELOPING HEALTHCARE FRAMEWORKS

Lata Virendrabhai Upadhyay

Assistant Professor, M.V.M College of Comm.,MGT & I.T., Saurashtra University

ABSTRACT

The domain of healthcare acquired its influence by the impact of big data since the data sources involved in the healthcare organizations are well known for their volume, heterogeneous complexity and high dynamism. Though the role of big data analytical techniques, platforms, tools are realized among various domains, their impact on healthcare organization for implementing and delivering novel use cases for potential healthcare applications shows promising research directions. In the context of big data, the success of healthcare applications solely depends on the underlying architecture and utilization of appropriate tools as evidenced in pioneering research attempts. Novel research works have been carried out for deriving application specific healthcare frameworks that offer diversified data analytical capabilities for handling sources of data ranging from electronic health records to medical images. In this paper, researcher presented various analytical avenues that exist in the patient centric healthcare system from the perspective of various stakeholders. Researcher also reviewed various big data frameworks with respect to underlying data sources, analytical capability and application areas. In addition, the implication of big data tools in developing healthcare eco system is also presented.

KEYWORDS: Big data, Healthcare, Framework, Infrastructure, Analytics

INTRODUCTION:

Big Data analytics and its implications received their own recognition in many verticals of which healthcare system emerges as one of the promising sectors (AndreuPerez et al., 2015). The distinguishing characteristics of big data namely Volume (hugeness of data availability), Velocity (arrival of data as a flood of fashion),

Variety (existence of data from multiple sources with diversified formats) find their own features in the abundant sources of healthcare data (MartinSanchez and Verspoor, 2014). The data sources for healthcare system have been broadly classified as (i) Structured data: Data that obeys defined data type, format, and structure. Example for such data in healthcare domain includes hierarchical terminologies of various diseases, their symptoms and diagnosis information, laboratory results, patient information such as admission histories, drug and billing information for the availed clinical services. (ii) Semi structured data: Data that has been organized with minimal structure along with self-describing nature. Example for such data includes data generated from devices such as sensors for effective monitoring of patient's behaviour. (iii) Unstructured data: Data that has no inherent structure, which may include medical prescriptions written in human languages, clinical letters, biomedical literature, discharge summaries and so forth. Hence, the exploration of healthcare data to achieve valuable insights for diversified stakeholders (clinicians, patients, hospitals, pharmacy etc.) is a challenging and daunting task due to the enormous variety of data (structured, unstructured, semi structured) from various sources. To extract Value (the fourth 'V' of big data attribute) from the existing 3 V's, it is appropriate to advocate efficient data processing platforms, smarter technologies for data collection, intelligent computational analysis, storage and visualization techniques (Sedg and Ola, 2014) towards attaining novel knowledge and efficient decision support strategies for different issues in healthcare.

Big data exhibits abundant potential to support a wide range of medical and healthcare functions such as clinical decision support, disease surveillance and population health management (Liang and Kelemen, 2016). Rapid advancement in Electronic Health Records (EHR) of patients, integration of social, behavioural and omics data with ICT based mHealth, eHealth, Smart Health and tele health devices have led to the development of novel healthcare frameworks for supporting precision medicine and personalized patient care. Recent research attempts show that the comprehensive healthcare solutions have resulted with the aid of architectural frameworks that supports various levels of layered services. The underlying idea behind the development of these frameworks is the effective analytics of various healthcare data sources for identifying casual relationships and pattern of interest among the sources of big data. These frameworks promote healthcare solutions from the disease centric model to patient centric model, where active participation of patients in their own healthcare resulted. In this paper, we review significant developments that occurred over recent years in building healthcare frameworks to orchestrate data analysis, management and visualization techniques as an end to end solution for the provision of improved healthcare services.

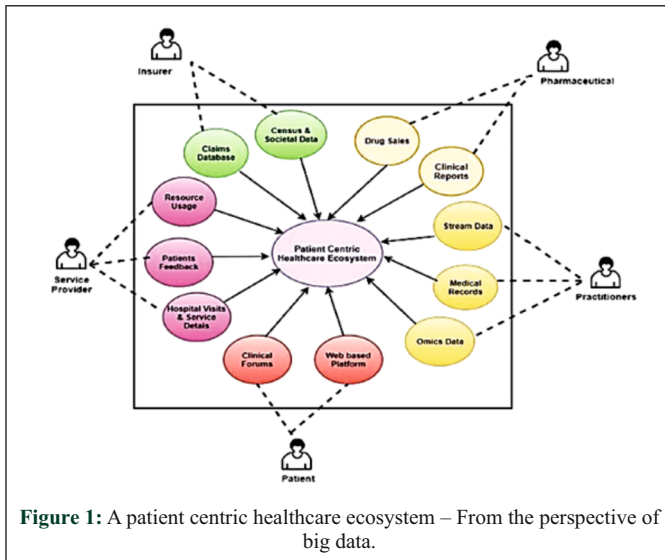
Due to the existence of diversified data formats, huge volume and associated uncertainty that exist among the sources of big data, the task of data curation plays a vital role in transforming raw data into an actionable knowledge. Though medical data are complex in nature, they exhibit strong interdependency and

hence tasks such as simplification of data complexity, identification of interconnection among various health features, selection of target attributes for healthcare analytics require highly sophisticated and matured domain specific tools and techniques. There are various diversified complexities that exist in the healthcare computing environment such as handling of stream oriented data from ubiquitous devices for the purpose of patient monitoring, integration of disparate data sources towards developing predictive models, format construction and compression of medical images. In addition, most of the data enrichment activities and knowledge synthesizing processes in the various stages of data analytics life cycle largely rely upon the usage of big data tools. In this perspective, the important features of various big data tools that play significant role in the construction of healthcare frameworks are highlighted by providing a systematic review.

2. THE IMPACT OF BIG DATA IN HEALTHCARE:

In the business sector, the core value of big data has been effectively utilized for the identification of behavioural patterns of the consumers to develop innovative business services and solutions. In the healthcare sector, the implication of big data serves predictive analytical techniques and machine learning platforms (AlJarrah et al., 2015) for the provision of sustainable solutions such as the implementation of treatment plans and personalized medical care. Jee and Kim (2013) compared the healthcare big data with the big data generated from the business sector under different attributes and their values. They redefined the characteristics of the healthcare big data into three features namely Silo, Security, and Variety instead of Volume, Velocity and Variety. Silo represents the legacy database that contains public healthcare information maintained in stakeholders' premises such as hospitals. The security feature implies the extra care needed in maintaining healthcare data. The variety feature indicates the existence of healthcare data in many forms such as structured, unstructured and semi structured.

With the advent of big data analytics and its associated technologies, the healthcare domain witnessed pragmatic transformations at various stages from the perspective of involved stakeholders (Wang and Alexander, 2015). The impact of big data in healthcare results in identifying new data sources such as social media platforms, telematics, wearable devices etc. in addition to the analysis of legacy sources that includes patient medical history, diagnostic and clinical trials data, drug effectiveness index etc. When the mixture of these data sources and analytics are coupled together, it provides a valuable source of information for healthcare researchers towards attaining novel healthcare solutions (Zhou et al., 2017). A typical patient centric healthcare ecosystem with its significant stakeholders and their diversified data sources (structured/semi structured/unstructured) is perceived in Fig. 1.



When these stakeholders work collaboratively and share their data insights effectively, healthcare solutions would be offered in a cost-effective manner with improved personalized care for patients.

By considering the importance of these stakeholders in building big data healthcare ecosystem, the next section throws insights on their perspective over the effective utilization of big data sources.

3. IMPLICATIONS OF BIG DATA TOOLS IN DEVELOPING HEALTHCARE FRAMEWORKS:

Though diversified big data frameworks are designed towards meeting specific healthcare objectives, they themselves orient well for adopting standard architectural guidelines for performing activities such as data gathering, preprocessing, data analysis, interpretation, and visualization. Due to domainspecific nature of big data healthcare framework, professionals such as data scientist should take utmost care in selecting appropriate tools (Philip Chen and Zhang, 2014; Sukumar et al., 2015) to be used at each level of the framework design and implementation. In this section, we insight the usage of various big data tools that play significant role in executing tasks such as integration of data, injecting intelligence, searching and indexing, stream data processing, and data visualization.

3.1 Data Integration Tools:

The continuous growth in the volume and velocity of healthcare data with diversified data types demands the necessity of utilizing the services of data integration tools for aggregating data from disparate sources.

Pentaho (2017) is a big data analytical platform that provides end-to-end data integration to support users for analyzing data from disparate sources such as relational databases, Hadoop distributions, NoSQL stores and enterprise applications. It also provides a flexible user interface for creating visual data flows to perform transformation and integration of data.

Palantir (2017) is a data integration tool that rapidly fuses data from disparate sources such as medical device outputs and medical codes. Further, it enables analytical techniques to develop models for tracking sequence of procedures and clinical data metrics to manage healthcare diagnosis. Ayata (2017) efficiently brings together structured and unstructured (videos, images, text, sound) healthcare data, mathematical models, business rules to build predictive and prescriptive models.

As data integration software, Attunity (2017) ingests data from disparate data sources in an efficient, cost effective and scalable fashion. It integrates data from major sources such as data warehouse, Hadoop and cloud platforms in a rapid manner without manual coding. Informatica (2017) offers a wide range of data management services that includes analytics, integration and governance of healthcare data for improved patient care. It accesses and transforms both clinical and administrative data that conforms to HIPAA (Health Insurance Portability and Accountability Act) and HL7 standards into usable format. Jitterbit (2017) offers a single secure platform for healthcare organizations to access clinical and workflow data by merging unstructured data with structured data in multiple standard formats. The data received from sources such as proprietary EHR systems can be transformed for comprehensive analysis.

3.2 Scalable Searching And Processing Tools:

Since large volume of clinical notes and unstructured text are commonly used by the physicians in the healthcare domain, there is an immense need for searching and indexing tools for performing optimized full text search capability of clinical data. These tools are utilized for effective distributed text management and

indexing large volume of data in file system such as HDFS (Hadoop Distributed File System).

Apache Lucene (2017) is a scalable, high performance indexing system that offers powerful and accurate full text search facility for variety of applications across different platforms. Google Dremel (Melnik et al., 2010) is a distributed system for interactively querying large data sets and supports nested data with column storage representation. It uses multilevel execution trees for query processing. Apache Drill is the Open Source implementation of Google Dremel.

Cloudera Impala (2017) offers high performance, low latency SQL queries on data stored in Apache Hadoop file formats. Impala integrates with the Apache Hive metastore database to share databases and tables between them. It pioneers the use of Parquet file format, a columnar storage layout optimized for large scale queries. Dryad is a general purpose distributed execution engine (Isard et al., 2007) that was designed to support a wide variety of parallel applications such as relational queries, large scale matrix computations and text processing tasks. It supports scalability by extending its processing capabilities from very small to large clusters. Application represented in the form of a data flow graph gets executed on a set of available computers, communicating through files, TCP pipes, and shared memory FIFOs.

3.3 Machine Learning Tools:

The healthcare industry is keen in availing the applications of machine learning tools to transform the abundant medical data into actionable knowledge by performing predictive and prescriptive analytics in view of supporting intelligent clinical activities.

Apache Mahout (2017) is an open source machine learning library that sits on top of Hadoop to facilitate the execution of scalable machine learning algorithms for a distributed environment. It offers techniques such as Recommendation, Classification, and Clustering. Mahout applications include pattern mining, user interest modeling, and personalization. Skytree (2017) is a general purpose machine learning platform that uses artificial intelligence to produce sophisticated algorithms for performing advanced analytics. It has the ability to process massive datasets (structured and unstructured) in an accurate manner without down sampling. Few of its use cases are recommendation systems, anomaly/outlier identification, predictive analytics, clustering and market segmentation, and similarity search.

Karmasphere (2017) creates a big data platform that mines and analyzes the web, mobile, sensor and social media in Hadoop. It provides a graphical environment that supports navigation through big data of any variety and spot trends and patterns in it. Karmasphere had been acquired by FICO in 2014. BigML, 2017 is a scalable and programmable machine learning platform that provides several tools to perform machine learning tasks such as classification, regression, cluster analysis, anomaly detection and association discovery. It seamlessly integrates the features of machine learning with cloud infrastructure to build cost effective applications with high scalability, flexibility, and reliability.

3.4 Real Time and Stream Data Processing Tools:

Advances in IoT and sensor devices found in healthcare domain prompts the data processing from diversified data sources to be carried out in a real time manner. The on the fly analysis of healthcare data enables the system to make better decisions for personalizing patient oriented services.

Apache Storm (2017) is a real time data platform for processing limitless streaming data that boasts capability to integrate seamlessly with existing queuing and database technologies to process over a million tuples per second per node. Its applications include real time analytics, interactive operation system, online machine learning, and ETL. S4 (Simple Scalable Streaming System) is a distributed stream processing engine (Neumeyer et al., 2010) that allows programmers to develop applications for processing continuous unbounded streams of data. Some of the key properties such as robustness, decentralization, scalability, cluster management, and extensibility have been offered by S4.

SQLstream Blaze (2017) is a streaming analytics platform that acquires data from all available sources in all formats and at all speeds (Unstructured, structured, local, distributed, in motion, at rest, live or historical). The Blaze has the potential to scale up its data handling capability to millions of records per second per CPU core with 1–5 ms latency. Splunk (2017) is a real time and intelligent big data platform that enables organizations to gain operational intelligence from machine data for real time insights. It performs indexing of structured/unstructured machine generated data, real time searching and reporting analytical results.

Apache Kafka (2017) is a distributed streaming platform used for building real time streaming data pipelines (reliable transfer of data between systems or applications) and applications (transform or act on streams of data). It uses four core APIs (Producer, Consumer, Streams, and Connector) to facilitate services such as message passing, storage and stream processing. SAP Hana (2017) is an in-memory analytics platform that provides real time analysis along with support for various capabilities such as database services, advanced analytics, application development, data access, administration, and openness.

3.5 Visual Data Analytical Tools:

Data visualization tools in healthcare helps to identify patterns, trends and deviations that include outliers, clusters, association discovery and time series analysis for improving clinical healthcare delivery and public health policy.

Jaspersoft (2017) is a scalable big data analytical platform that supports effective decision making with the help of interactive reports, analytics, and dashboards. It provides fast data visualization on storage platforms such as MongoDB, Cassandra, Redis, Riak, and CouchDB. Tableau (2017) has the ability to transform large, complex data sets into intuitive pictures. It combines advances in database and computer graphics technology to analyze huge datasets with limited computing resources. Qlik (2017) enables healthcare organizations to explore clinical, financial and operational data through visual analytics to discover insights that lead to improvements in care, reduced costs and delivering higher value to patients.

4. CONCLUSION:

Framework based solutions always cater to the comprehensive requirement of various stakeholders involved in the healthcare domain. With the impact of big data, healthcare domain was revamped and offer intensive solutions for handling diversified big data sources that range from patient health records to medical images. This paper reviews various research attempts in establishing healthcare frameworks and summarizes their significant outcomes. The summary of contributions by various researchers highlights the data source utilized, adopted analytical techniques and other features. At the end, the implication of various big data tools in developing healthcare framework is also extensively studied.

REFERENCES:

- I. AlJarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K., 2015. Efficient machine learning for big data: a review. *Big Data Res.* 2, 87–93. <https://doi.org/10.1016/j.bdr.2015.04.001>.
- II. AndreuPerez, J., Poon, C.C.Y., Merrifield, R.D., Wong, S.T.C., Yang, G.Z., 2015. Big Data for Health. *IEEE J. Biomed. Heal. Informat.* 19, 1193–1208. <https://doi.org/10.1109/JBHI.2015.2450362>.
- III. Bouhriz, M., Chaoui, H., 2015. Big data privacy in healthcare moroccan context.
- IV. Chawla, N.V., Davis, D.A., 2013. Bringing big data to personalized healthcare: a patientcentered framework. *J. Gen. Intern. Med.* 28, 660–665. <https://doi.org/10.1007/s1160601324558>.
- V. Dencelin, L.X., Ramkumar, T., 2016. Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures. *Biomed. Res.* 2016, S166–S173.
- VI. Dimitrov, D.V., 2016. Medical internet of things and big data in healthcare. *Healthc. Inform. Res.* 22, 156–163. <https://doi.org/10.4258/hir.2016.22.3.156>.
- VII. Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D., 2007. Dryad: distributed dataparallel programs from sequential building blocks. *ACM SIGOPS Oper. Syst. Rev.* 59–72. <https://doi.org/10.1145/1272998.1273005>.
- VIII. LegazGarcia, M.D.C., MartinezCosta, C., MenarguezTortosa, M., FernandezBreis, J. T., 2016. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. *Knowledge Based Syst.* 105, 175–189. <https://doi.org/10.1016/j.knosys.2016.05.016>.
- IX. Liang, Y., Kelemen, A., 2016. Big data science and its applications in health and medical research: challenges and opportunities. *J. Biom. Biostat.* 7, 1–9. <https://doi.org/10.4172/21556180.1000307>.
- X. Lin, W., Dou, W., Zhou, Z., Liu, C., 2015. A cloud based framework for Homediagnosis service over big medical data. *J. Syst. Softw.* 102, 192–206. <https://doi.org/10.1016/j.jss.2014.05.068>.
- XI. Mahmud, S., Iqbal, R., Doctor, F., 2016. Cloud enabled data analytics and visualization framework for health shocks prediction. *Futur. Gener. Comput. Syst.* 65, 169–181. <https://doi.org/10.1016/j.future.2015.10.014>.
- XII. Mancini, M., 2014. Exploiting big data for improving healthcare services. *J. ELearning Knowl. Soc.* 10, 23–33.
- XIII. MartinSanchez, F., Verspoor, K., 2014. Big data in medicine is driving big changes. *Yearb. Med. Inform.* 9, 14–20. <https://doi.org/10.15265/IY20140020>.
- XIV. Pramanik, M.I., Lau, R.Y.K., Demirkan, H., Azad, M.A.K., 2017. Smart health: big data enabled health paradigm within smart cities. *Expert Syst. Appl.* 87, 370–383. <https://doi.org/10.1016/j.eswa.2017.06.027>.
- XV. Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Heal. Inf. Sci. Syst.* 2, 3. <https://doi.org/10.1186/2047250123>.
- XVI. Xavier, L., Thirunavukarasu, R., 2017. A distributed treebased ensemble learning approach for efficient structure prediction of protein. *Int. J. Intell. Eng. Syst.* 10, 226–234. <https://doi.org/10.22266/ijies.2017.0630.25>.
- XVII. Youssef, A.E., 2014. A Framework for secure healthcare systems based on big data analytics in mobile cloud computing environments. *Int. J. Ambient Syst. Appl.* 2, 1–11. <https://doi.org/10.5121/ijasa.2014.2201>.
- XVIII. Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine learning on big data: opportunities and challenges. *Neurocomputing* 237, 350–361. <https://doi.org/10.1016/j.neucom.2017.01.026>.