The Impact of Big Data In Healthcare Analytics

Usman Akhtar
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
usman@oslab.khu.ac.kr

Jong Won Lee
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
jongwlee@khu.ac.kr

Hafiz Syed Muhammad Bilal
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
bilalrizvi@oslab.khu.ac.kr

Taqdir Ali
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
taqdir.ali@oslab.khu.ac.kr

Wajahat Ali Khan
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
wajahat.alikhan@oslab.khu.ac.kr

Sungyoung Lee
Department of Computer Science & Engineering, Kyung Hee University,
Yongin-si, 17104, South Korea
sylee@oslab.khu.ac.kr

***Abstract*—No doubt that we are living in the era of Big Data, where we are noticing the expansion of smart healthcare devices. The main obstacles for the Healthcare platform researchers in choosing the right Big Data tool to process unstructured data. Therefore, the current area of research is shifted from massive storage to efficiently analyze the data. This paper aims to present state-of-the-art Big Data analytics tools and presented the Intelligent Medical Platform (IMP) as a case study in dealing with the multimodal data. The result shows that the proposed platform is scalable in dealing with health care data.**

***Index Terms*—multimodal sensory data, Big Data, data acquisition and persistence**

I. Introduction

In the recent past, a tremendous amount of data is being produced at an alarming rate in all medical data centers. The volume of data is predicted to reach 35 zettabytes by 2020 [1]. Therefore, maintenance and processing of large scale data is creating a challenge for the cloud data centers. Thus, effective processing of cloud storage is required for health and wellness platforms. Currently, the research trends is shifted from massive storage to effective processing of healthcare data. This change has made an impact on the conventional data platform and models. With the advent of smart devices, the opportunity has emerged in the field of the medical platform to empowers the user by providing effective and timely data analytics support.

Big data in healthcare are concerned with the complex data that are too big for state-of-the-art solutions to perform healthcare analytics. The key attributes of Big Data include **Volume** (the rate at which the data is growing), **Velocity** (arrival of data), **Variety** (heterogeneous format of data such as structured and unstructured data) and **Value** (that is the ability to transform the data into a meaningful way). In healthcare, the data is acquired from heterogeneous sources such as clinical, health and organizational records. Therefore, designing a scalable Big Data analytics system has to face challenges as processing these heterogeneous data is difficult and these data need to store in realtime while maintaining the performance guarantee. Following are the main challenges:

TABLE I
Machine Learning tools for big data

Tools	***Features***	***OpenSource***
Apache Mahout	Offers a scalable machine leaning algorithms for clustering and classification	✓
Skytree	Support the mining of massive datasets and povide analytics in real time	⊗
Karmasphere	Support for Big Data tool fro pattern mining on large scala Hadoop cluster	⊗
Jaspersoft	Provide an interactive analytics support and data integration from diver sources	✓

- The growth of the healthcare record has led to greater use of the Clinical Decision Support System (CDSS) in healthcare informatics. For streaming data in healthcare the existing traditional database systems are not enough to handle the medical data.
- Providing a reliable services with high availability are the essential requirements that need to be fulfilled. Currently, healthcare analytic systems use the computationally expensive services that slow down the performance query time.
- To detect the warning sings at realtime required a stream computing platform. The current stream processing platform offers very limited support. Therefore, for the better performance customize solution of the Apache Kafka[1] and Apache Storm[2] is required to fulfill the diverse healthcare scenarios.

A lot of efforts have been made to deal with the realtime analytics of the health records. The diversified framework is currently designed and use to meet the Big Data analytics requirements to gather, process and analyze the data. The healthcare industry now looking for solutions in machine

[1] https://kafka.apache.org/

[2] https://storm.apache.org/

TABLE II
SUMMARY OF THE STATE-OF-THE-ART APPROACHES FOR BIG DATA ANALTICS.

Work	Features	Category	Data Sources	Analytics Capability	Application
Raghupathi et al. [2]	Perform the extensive support for the big data analytics	No prototype available only generic system	Support heterogeneous data	Queries	Recommendation systems
Chawla et al. [3]	Provide the data driven support to perform the Big Data analytics	Patient centric platform	medical records, EMR, and clinical notes	Support for filtering	Personalized healthcare system
Kim et al. [4]	Support for Big Data platform for vital signs	Data analytics and Visualiazation framework	medical records and ECG	Querying over the Big Data Platform	Personalized services for healthcare
Fang et al. [5]	Support for Big Data to perform the health informatics	Conceptual framework	Patient profiles and clinical data	Querying	Healthcare services for Big Data
Youssef et al. [6]	Support for Big Data analytics in mobile computing environment	Healtcare information system	EHR and critical signs	Big data quering using MapReduce	Services for smart healthcare systems
Lin et al. [7]	A cloud based system for healthcare analytics	Analytical services for cloud	Clinical records	MapReduce	Smart system technologies for hospitals
Sakr and Elgammal [8]	Provide data integration support	Conceptual framework	Patient profile data	MapReduce and SPARK streaming	Descriptive analytics

learning to applied to the Big Data storage platform. The purpose of the machine learning tool is to transform the unstructured data into actionable knowledge as shown in Table I. Apache Mahout [2] is an open source machine learning tool to apply the classic algorithms on unstructured data. Currently, Apache Mahout is utilized in the recommendation system, CDSS, and expert system. Skytree [2] is another machine learning platform for Big Data that utilized Artificial Intelligence (AI) to produce advance data analytics on unstructured data. Karmasphere [2] is another platform design to mine the social web for extracting the knowledge. This platform uses Apache Hadoop for its storage and includes many Machine Learning tasks such as classification, regression and anomaly detection. Jaspersoft [2] provides an interactive dashboard to perform Big Data analytics and visualization. The goal of our work is to figure out the key challenges in the existing state-of-the-art solutions to handle Big Data streams and provide analytics. As real-time analytics is the key requirement for healthcare systems.

In this paper, we studied the state-of-the-art data Big Data analytics platform for healthcare and highlights the contribution in the field of healthcare. We also introduced the realtime scenario as a case study for dealing with the volume and variety of the medical data.

II. STATE-OF-THE-ART BIG DATA FRAMEWORKS FOR HEALTHCARE

In recent years, various platforms are reported to deal with the new requirements of healthcare. The Big Data analytics framework proposed by Raghupathi et al. [2] that consists of the Data Source layer, Data Transformation layer and Big Data storage for analytics. The role of the data source layer is the acquisition of multiple data sources from diverse medical devices in various formats. The ETL operations are performed by the Data Source Layer and the Big Data Layer consist of the Hadoop Ecosystems and tools that are responsible for querying over the distributed stored data. However, the major limitation of the work is that they only perform the qualitative comparison with the other approaches and no realtime implementation is performed. Chawla et al. [3] proposed a data analytics system that is based on collaborative filtering to predict user opinion and individual patient health history. The major benefit is to predict disease based on similarity computation. Kim et al. [4] proposed a system that provides healthcare services based on the analysis of the vital signs. It utilized the Apache Hadoop based platform for the extraction of the vital signs. Fang et al. [5] surveyed computational aspects of health informatics to extract the meaningful patterns from the Big Data also highlights the technical aspects of the Big Data tools. With the advent of mobile devices, the new solutions are reported in the literature. Youssef et al. [6] proposed a framework for healthcare Big Data analytics in the mobile computing environment. Their framework provides a high-level integration of healthcare among different EHR components to share a common format in the different healthcare organizations. The need for the smart services for patients are on a high demand Lin et al. [7] proposed a Big Data analytics systems that expedite the overall query processing time in a Hadoop cluster. A system for the healthcare framework proposed by Sakr and Elgammal [8] that integrate the healthcare technologies and provide a data connection, data storage and Big Data processing layer to perform the descriptive analytics.

III. CASE STUDY: MEDICAL DATA PERSISTENCE

The advances in the healthcare industry have witnessed great progress in generating the data from the smart medical processing. These advancements have made tremendous growth in generating healthcare data, but also make it difficult to handle and process. Therefore, we have proposed Big Data *Intelligent Medical Platform (IMP)*[3] to efficiently acquire, synchronize and manage medical Big Data from diverse sources.

[3] http://imprc.cafe24.com/

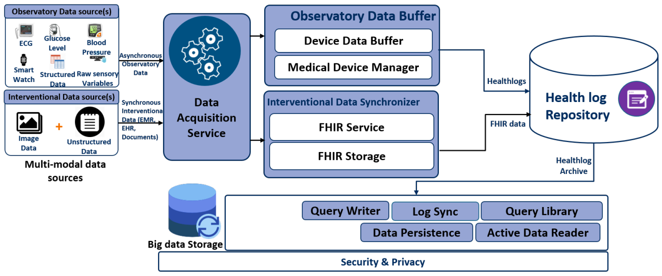
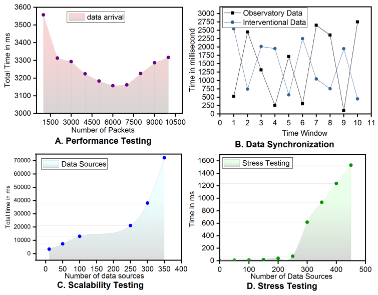


Fig. 1. Showing the Big Data Persistence of IMP platform.

To provide a more convenient service and environment of healthcare, IMP provides a patient-centric healthcare application and services. Considering the health data as an asset Data Acquisition and Persistence layer acquires the multimodal and legacy system data to be securely stored into a Big Data infrastructure. The asset of the IMP platform is the storage of the multimodal data sources. These data that are coming from the devices that are acquired and stored over the non-volatile distributed storage; which is primary build over the Hadoop on a private cloud. Diverse formats of data are currently being stored in the IMP storage such as multimodal data sources, FHIR based interoperable data and Archive data. The main component of IMP is non-volatile data persistence. This component used Apache Hive queries to extract data from the non-volatile storage. To access the data, the IMP support data reader is used to provide a continuous response in real-time for data analytics and visualization. The current implementation of the data reader in IMP support *Apache Hive*[4] based predefined queries that match the request from the client for a stream-based response. As shown in the Fig. 1, the observatory and interventional data are received by the data acquisition service and data buffer temporary holds the buffer before sending to the Health log repository and the archived are stored inside the Big Data storage for non-volatile persistence.

IV. Performance Evaluation

We have evaluated the performance of the IMP data acquisition and persistence services as shown in Fig. 2. In the performance testing as shown in Fig. 2A multiple requests from the user are generated and synchronized at the server. The accuracy of the data acquired is measured 1500 packets containing the sensory data are successfully synchronized. In Fig. 2B, we have increased the time window to get the data from the observatory and interventional data sources, the result shows the accuracy of the data acquired from both the sources. To scale the IMP platform with the increase in the number of multimodal data sources we have incrementally increased the number of the data sources from 50 to 400 with the fixed packet size as shown in Fig. 2C. The result shows that our implementation was able to scale successfully even the number of the devices were increased. To check the effectiveness of the IMP data acquisition and persistence platform we performed stress testing with 500 data sources as several times incrementally increased with the number of devices as shown in Fig. 2D.



A. Performance Testing

B. Data Synchronization

C. Scalability Testing

D. Stress Testing

Fig. 2. Performance evaluation of the IMP platform for Big Data Analytics.

V. Conclusion

In this paper, we presented the implication of the Big Data in creating a Healthcare framework. We highlighted the shortcoming of the existing Big Data analytics tools in dealing with the evolution of data. The proposed IMP Big Data storage is a promising solution for dealing the heterogeneous health data. The result shows the effectiveness of the proposed platform in terms of better query performance and scalability in distributed systems. In future, we will compare the scalability of the proposed framework with the other platform.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion)" and IITP-2017-0-00655 and NRF-2016K1A3A7A03951968 and NRF-2019R1A2C2090504 and NRF-2019R1G1A1011296

References

[1] Z. Goli-Malekabadi, M. Sargolzaei-Javan, and M. K. Akbari, "An effective model for store and retrieve big health data in cloud computing," *Computer methods and programs in biomedicine*, vol. 132, pp. 75–82, 2016.

[2] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, p. 3, 2014.

[3] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of general internal medicine*, vol. 28, no. 3, pp. 660–665, 2013.

[4] T. W. Kim, K. H. Park, S. H. Yi, and H. C. Kim, "A big data framework for u-healthcare systems utilizing vital signs," in *2014 International Symposium on Computer, Consumer and Control*. IEEE, 2014, pp. 494–497.

[5] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar, "Computational health informatics in the big data age: a survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 12, 2016.

[6] A. E. Youssef, "A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments," *Int J Ambient Syst Appl*, vol. 2, no. 2, pp. 1–11, 2014.

[7] W. Lin, W. Dou, Z. Zhou, and C. Liu, "A cloud-based framework for home-diagnosis service over big medical data," *Journal of Systems and Software*, vol. 102, pp. 192–206, 2015.

[8] V. Palanisamy and R. Thirunavukarasu, "Computer and information sciences," 2017.

[4] https://hive.apache.org/