# Challenges in Managing Real-Time Data in Health Information System (HIS)

Usman Akhtar[1], Asad Masood Khattak[2], and Sungyoung Lee[1(✉)]

[1] Department of Computer Science and Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, South Korea
{usman,sylee}@oslab.khu.ac.kr
[2] College of Techological Innovation, Zayed University, Dubai, UAE
asad.khattak@zu.ac.ae

**Abstract.** In this paper, we have discussed the challenges in handling real-time medical big data collection and storage in health information system (HIS). Based on challenges, we have proposed a model for real-time analysis of medical big data. We exemplify the approach through Spark Streaming and Apache Kafka using the processing of health big data Stream. Apache Kafka works very well in transporting data among different systems such as relational databases, Apache Hadoop and non-relational databases. However, Apache Kafka lacks analyzing the stream, Spark Streaming framework has the capability to perform some operations on the stream. We have identified the challenges in current real-time systems and proposed our solution to cope with the medical big data streams.

**Keywords:** Stream processing framework · Health-care Information System (HIS) · Kafka messaging

## 1 Introduction

Over the past two decades, the technology advancements have led to increase data from many domains like health care and other scientific sensors. Big data in health is concerned with the datasets that are too big and complex to process and interpret with the existing tools and these datasets present a problem with the storage, analysis and visualization. Big data is unstructured and normally require real-time analysis. Big data refers to those datasets that are very large and complex to manage with traditional software or common data management tools. Big data in health-care is growing enormous not only because of its volume but also with the diversity of the datatypes. In health-care sector, there are three main types of digital data: clinical records, health research records and organization operations records. Clinical records consist of a variety of data like electronic medical records, images and wireless medical devices and these devices are the major contributors to the flood of big data. Health research records also generate large amount of data that are unstructured. Genetic differences

© Springer International Publishing Switzerland 2016
C.K. Chang et al. (Eds.): ICOST 2016, LNCS 9677, pp. 305–313, 2016.
DOI: 10.1007/978-3-319-39601-9_27

study, which include more than 100,000 the participants, generate approximately 1.5 GB of data per person. Organization operations records such as billing and scheduling have been digitized resulting in large amount of data generation.

Designing a scalable big data system faces a series of challenges. First, due to the heterogeneous and huge volume of data, it is difficult to collect data from the distributed locations. Second, storage is the main problem for heterogeneous datasets. Big data system need to store while providing performance guarantee. Third challenge is related to mining massive datasets at real-time that include visualization. These challenges demand a new processing paradigm as the current data management systems is not efficient in dealing with the real-time or heterogeneous nature of data. However, traditional database management system is based on storing the structured data in relational database management system (RDBMS). These traditional systems do not provide any support for unstructured or semi-structured data. From the view of scalability, there are many flaws in traditional RDBMS in scaling for managing with the hardware in parallel, which is not suitable to manage growing data. To address these challenges, many solutions have been proposed by the research community such as NoSQL [1], which is more suitable for dealing with the massive heterogeneous data. MapReduce [2] programming model is more suitable for processing large datasets as it consists of map and reduce function. Apache Hadoop is a batch processing system that allows distributed processing of large datasets. It also integrates storage, data processing and other management modules to provide a powerful solution. One of the major limitation of Hadoop is dealing with the real-time stream processing or in memory processing.

Real-time management of data stream and providing data analytics is the key requirement in health care [3]. Data stream refer to those data that continuously arrive without persistent in the storage. Fast inference in real-time is still a major research problem. The streaming processing paradigm will normally analyze data as soon as it arrives and data is in the form of stream. Only a small portion of the stream is stored in limited memory. Some of the distributed stream computing framework include Storm, S4, Kafka and Storm Streaming [4].

Research on health-care management is constantly growing. Research community is interested in storing data captured from medical devices. Managing stream is an ongoing research issue and much of the research work is related to the querying stream data and management. Data storage is one of the crucial part of the health-care information system as the data are coming from the different sources, PolyglotHIS [5] combines relational, graph and other document models to conciliate variety of information. In the recent past, there has been a growing attempt to utilize medical signal analytics to improve patient care management [6].

### 1.1 Motivation

Research on real-time health-care data acquisition and management is constantly growing. The main motivation is to figure out the challenges in health information system (HIS), while dealing with the real-time data stream. Moreover, we

**Table 1.** Real-time systems spectrum

| Current focus | Batch processing | Relational database management system |
|---|---|---|
| Technologies | HDFS, MapReduce | MySQL, PostgreSQL |
| Limitations | Suitable for batch processing not for the real-time streams | Not suitable for unstructured data |
| Challenges | Continuous processing of stream, load balancing | Not suitable for generating real-time analytics |
| Possible solution | Use KafKa for distributed stream processing and Apache Spark streaming to perform operations on the streams | Use of NoSQL, Apache Hive and ApacheHbase for both the structured and unstructured data |

have worked out to find the limitations of existing technologies. Our main goal is to develop a model that is capable of handling medical big data streams and provide analytics.

Table 1, showing the current real-time system spectrum as shown, the current focus is mainly on the batch processing and using the relational database management system. For managing real-time medical streams batch processing is not the appropriate solution. Possible solution is the uses of distributed messaging system like Apache KafKa and store the data in NoSQL databases for analytics.

We have discussed the challenges for the health information system while dealing with the big data in Sect. 2. Based on the challenges discuss in this paper we have proposed our model in Sect. 3.

## 2 Big Data: Challenges in Health Information System

Health-care industry is generating the massive amount of data. To analyze and store real-time management of medical stream is a challenging task. We have identity the challenges in health information system and briefly describe the current technology available to cope with these challenges. We also consider the limitation of the current technology.

### 2.1 Challenges in Health-Care Informatics

The speedy growth of health-care organization and also the continuously increasing number of patients has led to the greater use of the clinical support system. Modern medical imaging techniques are capable of producing high resolution images but they require High Performance Computing. Many monitoring devices continuously generate data and the data is normally stored for a small period of

time. However, many attempt have been made to utilize the data from telemetry and other physiological monitoring devices. Many areas in health-care can be improved by the use of computational intelligence. In variety of medical applications different image acquisition techniques have been utilized.

For health-care informatics challenge current technology like HBase and MongoDB are becoming more common in research communities. Apache Hadoop is highly scalable and open source framework which render variety of computing modules such as Spark Streaming and Apache Storm [7]. For performing variety of operations while dealing with the medical streaming data spark and storm is promising platform. These platforms are capable of ingesting and computing streaming data. In intensive care unit (ICU), vast amount of data is produced and there is a need to develop clinical decision support system (CDSS) [8].

For the streaming data in health-care, there is a need to develop a platform as the current technologies are limited to handle medical devices data. Moreover, data acquisition and ingestion is required for clinical health-care setting, while dealing with the streaming data. The main challenges while dealing with the data acquisition are network bandwidth, scalability and cost issues.

### 2.2 Increased Cost with Low Health-Care Services

There are plenty of challenges that health-care are facing specially hospital administrator and researchers. However, consumers are experiencing increased cost without any gain in health services.

Currently, health-care application uses very expensive system to manage data. The health-care system should use Hadoop, as it is an open source software from Apache and it include MapReduce framework for running applications. Hadoop is very scalable and supports many other modules for distributed storage and computations. Hadoop also provides many high language support like Pig Latin [9] and Hive [10] and many vendors are offering commercial distributions including Cloudera, MapR, and Hortonworks.

### 2.3 Stream Computing for Smarter Health-Care

To improve the patient outcome and to detect the warning signs of the complications, there is a need for stream computing platform. In modern health-care patients are routinely connected to the medical equipment that continuously monitor blood pressure and heartbeat.

Consequently, data processing engines are gaining popularity in health-care such as Spark Streaming, Apache Storm and Apache Kafka. Apache Kafka is mainly used for stream processing and it supports wide range of the scenarios and claims high throughput and reliability. Apache Storm processes data in real-time and it is extremely useful in processing high velocity of data. While Apache Spark is in-memory data processing and includes rich APIs support to efficiently execute streaming and allow SQL for iterative access of data. While dealing with the fast moving big data analytics, Apache Storm and Apache Kafka complement each other. Most stream distributed processing systems, like Apache Storm, work

on a record at a time model, where each node is responsible for receiving the data and update internal state.

These models raise many challenges in cloud environment in term of fault tolerance, consistency and fusion with the batch processing. A new programming model was proposed which consists of discretized streams (D-Streams). Apache Spark model uses small intervals and treats the stream a series of deterministic batch processing computations. Apache Spark is efficient in building a modern health care application that fuses well with the batch processing. While dealing with the streaming data, Apache Spark is fault tolerant as it employs a new approach called parallel recovery.

### 2.4 Transforming Health-Care Data into Information

In the recent advancements there are lot of Real-Time technologies like Kafka, Storm, and Spark Streaming, but the main problem is the lack of knowhow in using these technologies.

**The Stream Data Platform:** The stream data processing is different than the query processing in traditional relational database management system. Our main goal is to the design a system that stores the patients physiologic features and displays the analytics in real-time. These analytics help the medical staff to initiate medical intervention earlier to save lives.

But there are some challenges while dealing with the stream data platform. The first problem in building the health-care application is the transportation of the data between systems. A lot of data systems are involved in building some health-care applications, such as, relational databases, Hadoop, search system and other analytics system. As in the distributed system all of these need reliable feeds of data and this raises the problem of data integration when you have to cope with the unstructured data.

Second problem is related to the analytical data processing e.g. processing in data warehouse. The inefficient data integration between the system makes it hard to perform richer analytical data processing.

**Stream Data Collection:** In a typical hospital, nurses are responsible for monitoring patients vital signs manually. They visit each patient to record the critical signs but the condition of the patient may decline after the visit. Currently, new wireless sensors are available that can transmit the patient conditions at much higher frequencies and these measurements can be streamed to Hadoop for storage. In ICU, patient monitoring system monitor the vital signs such as heart rate, respiration rate, body temperature and blood pressure. Data stream of these vital signs can be sent via a communication port supported by the interface software and sequentially stored in Hadoop for distributed data storage.

## 3 Discussion

Based on the challenges discuss in the health-care application we have proposed a solution based on the KaKfa data streaming and Spark Streaming. We believe that this model can work indecently and with the existing application. Efficient processing of data in health-care increases the quality of patient monitoring. Consider, the Patient with a Cardiac problem, and often wear Cardiac Event Recorders (CERS), which constantly record the Hearts Electrical Activity (ECG). In CERS data is manually transferred to other device. So, this stream can be utilized to perform real-time analysis. Medical devices are the ultimate source of sending the data. But these data cannot be sent directly to the HDFS for storage and analysis. Kafka is used to buffer the data before writing into the HDFS. Frameworks are used to perform some operations on the stream, these frameworks are Spark Streaming and Apache Storm. To query the historical data in HDFS, Apache Hive and Apache Impala are widely used Fig. 1. Shows data sources coming from the medical devices to Kafka as it provides the functionality of the messaging system. To perform operations on the stream, Spark Streaming is applied. This will process the stream and finally write it on the HDFS. Hive runs on top of the Hadoop to query the data inside the HDFS for visualization and analytics.

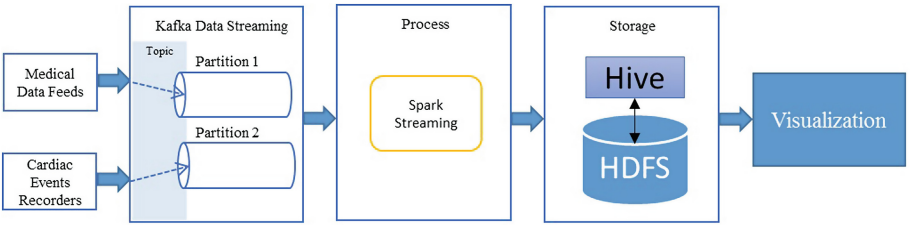


**Fig. 1.** Transforming health-care data into information

Kafka data streaming module is responsible for managing the events streams. A stream that is coming in the Kafka is modeled by a topic which gives name to the data. As Kafka is a publish-subscriber messaging system each message has its own key, which is useful way to partition the data on cloud. In medical big data application ETL (extract, transform and load) pipelines are needed as data platform is being able to stream data between systems to provide analytics.

## 4 Related Work

Big Data application are widely used in health-care, medical and government services. Recently, national health authority of Australia encloses Personally Controlled Electronics Health Record (PCEHR) to manage the individual health

records as an electronics health records (eHR) [11]. The volume of these health-care systems are very large and it is very useful for the doctors and other health-care providers to provide the best possible care. Ehealth system are gaining lot of importance in research community and researchers put emphasis on the privacy, data integrity and storage. Some commercial companies are trying to propose there own solution to deals with the stream processing like IBM InfoSphere Streams which has the capability to analyze data in motion and can easily integrate with the development framework like Eclipse and enable you to analyze, visualize and test with the help of Streams Processing Language (SPL) [12].

In Table 2, we have shown the comparison of different eHealth frameworks proposed in the literature. Now researchers are trying to propose a model that include Big Data ecosystem technologies in more effective way. Health-care need real-time data analysis, we have taken a step to propose a model that uses Big Data processing and analytics for Health-care scenarios.

**Table 2.** Comparison of different eHealth frameworks

| eHealth proposed models | Big data system | Real-time analysis | Privacy and security | Limitation |
|---|---|---|---|---|
| A privacy-preserving framework for personally controlled electronic health record (PCEHR) system [13] | No | No | Yes | Major focus is on the privacy, storage of large health record is still the main challenge |
| An open platform for personal health record apps with platform-level privacy protection [14] | No | No | Yes | Proposed a cloud based solution but uses relational database for storage. Not suitable for real-time |
| Privacy preservation and information security protection for patients portable electronic health records [15] | No | No | Yes | Useful for safeguarding the EHR, but storage and visualization is not covered |
| The Taiwanese method for providing patients data from multiple hospital EHR systems [16] | Yes | No | Yes | Useful for safeguarding the EHR, but storage and visualization is not covered |

## 5 Conclusion and Future Work

To improve the patient outcome and to detect the warning signs of the complications, there is a need of a streaming computing platform. The traditional relational system does not provide any support for unstructured nature of medical data stream. From the view of scalability, there are many flaws in traditional RDBMS in scaling with the hardware in parallel, which is not suitable to manage growing data. After analyzing the challenges we have proposed a cheap commodity health-care information system which can successfully process real-time streaming and store data on Hive warehouse on the top of the HDFS. The main limitation of the work is not able to perform end-to-end evaluation of the work along with the response time. In future, we will continue to work on how we can efficiently capture data from the medical devices regardless of the data format.

**Acknowledgments.** This work was supported by the Industrial Core Technology Development Program (10049079, Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) and This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2011-0030079). This research work was also supported by Zayed University Research Initiative Fund R15098.

## References

1. Cattell, R.: Scalable SQL, NoSQL data stores. SIGMOD Rec. **39**(4), 12–27 (2010). http://doi.acm.org/10.1145/1978915.1978919
2. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008). http://doi.acm.org/10.1145/1327452.1327492
3. Peek, N., Holmes, J., Sun, J.: Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. Yearb Med Inform **9**(1), 42–7 (2014)
4. Zaharia, M., Das, T., Li, H., Shenker, S., Stoica, I.: Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters. In: Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing, HotCloud 2012, Berkeley, CA, USA, p. 10. USENIX Association (2012). http://dl.acm.org/citation.cfm?id=2342763.2342773
5. Kaur, K., Rani, R.: Managing data in healthcare information systems: many models, one solution. Computer **3**, 52–59 (2015)
6. Apiletti, D., Baralis, E., Bruno, G., Cerquitelli, T.: Real-time analysis of physiological data to support medical applications. Trans. Info. Tech. Biomed. **13**(3), 313–321 (2009). http://dx.doi.org/10.1109/TITB.2008.2010702
7. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 1–10 (2014). http://dx.doi.org/10.1186/2047-2501-2-3

8. Hussain, M., Khattak, A., Khan, W., Fatima, I., Amin, M., Pervez, Z., Batool, R., Saleem, M., Afzal, M., Faheem, M., et al.: Cloud-based smart cdss for chronic diseases. Health Technol. **3**(2), 153–175 (2013)
9. Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A.: Pig latin: a not-so-foreign language for data processing. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, pp. 1099–1110. ACM, New York (2008). http://doi.acm.org/10.1145/1376616.1376726
10. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., Murthy, R.: Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. **2**(2), 1626–1629 (2009). http://dx.doi.org/10.14778/1687553.1687609
11. Rabbi, K., Kaosar, M., Islam, M.R., Mamun, Q.: A secure real time data processing framework for personally controlled electronic health record (PCEHR) system. In: Tian, J., Jing, J., Srivatsa, M. (eds.) SecureComm 2014, pp. 141–156. Springer, Heidelberg (2014)
12. Nabi, Z., Wagle, R., Bouillet, E.: The best of two worlds: integrating IBM infosphere streams with apache YARN. In: 2014 IEEE International Conference on in Big Data (Big Data), pp. 47–51. IEEE, (2014)
13. Begum, M., Mamun, Q., Kaosar, M.: A privacy-preserving framework for personally controlled electronic health record (PCEHR) system (2013)
14. Van Gorp, P., Comuzzi, M., Jahnen, A., Kaymak, U., Middleton, B.: An open platform for personal health record apps with platform-level privacy protection. Comput. Biol. Med. **51**, 14–23 (2014)
15. Huang, L.-C., Chu, H.-C., Lien, C.-Y., Hsiao, C.-H., Kao, T.: Privacy preservation and information security protection for patients portable electronic health records. Comput. Biol. Med. **39**(9), 743–750 (2009)
16. Jian, W.-S., Wen, H.-C., Scholl, J., Shabbir, S.A., Lee, P., Hsu, C.-Y., Li, Y.-C.: The taiwanese method for providing patients data from multiple hospital EHR systems. J. Biomed. Inform. **44**(2), 326–332 (2011)