2018 IEEE 4th International Conference on Collaboration and Internet Computing

Assessment of an IoT Platform for Data Collection and Analysis for Medical Sensors

1st João Rei
Escola de Engenharia
Universidade do Minho
Braga, Portugal
a68381@alunos.uminho.pt

2nd Cláudia Brito
Escola de Engenharia
Universidade do Minho
Braga, Portugal
a71575@alunos.uminho.pt

3rd António Sousa
Departamento de Informática
Universidade do Minho
Braga, Portugal
als@di.uminho.pt

***Abstract*—Health facilities produce an increasing and vast amount of data that must be efficiently analyzed. New approaches for healthcare monitoring are being developed every day and the Internet of Things (IoT) came to fill the still existing void on real-time monitoring. A new generation of mechanisms and techniques are being used to facilitate the practice of medicine, promoting faster diagnosis and prevention of diseases. We proposed a system that relies on IoT for storing and monitoring medical sensors data with analytic capabilities. To this end, we chose two approaches for storing this data which were thoroughly evaluated. Apache HBase presents a higher rate of data ingestion, when collaborating with the Kaa IoT platform, than Apache Cassandra, exhibiting good performance storing unstructured data, as presented in a healthcare environment. The outcome of this system has shown the possibility of a large number of medical sensors being simultaneously connected to the same platform (6000 records sent by the second or 48 ECG sensors with a frequency of 125Hz). The results presented in this paper are promising and should be further investigated as a comprehensive system would benefit the patient's diagnosis but also the physicians.**

***Keywords*-Big Data, healthcare, IoT, sensors, Apache, HBase, Cassandra, KaaIoT, remote monitoring**

I. Introduction

Nowadays, healthcare organizations store their medical records in electronic databases and data transparency is being applied to make that stored data usable, searchable and actionable [1]. As healthcare data is rapidly increasing, healthcare providers and data scientists now have access to promising new threads of knowledge, called "big data" as result of not only its volume but also its complexity and diversity [1], [2]. The analysis of this data, in order to obtain useful insight, allows healthcare providers to deliver higher quality care and reduce costs. [1], [3]. Many innovative companies in the private sector are building applications and analytic tools that help patients, physicians and other healthcare stakeholders to identify value and new opportunities based on all this emerging data [1], [3].

However, there are some obstacles in healthcare systems that big data has to overcome in order to succeed [4], [5]. The privacy of the data is of the most important aspects as more information becomes public and big data systems need to be carefully developed to satisfy storage and analytic requirements [1], [4], [6].

Thus, new data management systems have been developed to face these challenges such as Hadoop [7], an open source framework that helps solve problems related to storage and access to data and allows very fast parallel processing; HBase [8], an open source distributed database that is built on top of Hadoop's file system; and Cassandra database [9], that can store millions of columns in a single row and does not require prior knowledge of data formatting [6], [10]. Additionally, for big data analytic purposes, Spark [11] is one of the most widely used open source processing frameworks [12].

One of the major sources of big data is the Internet of Things (IoT). This concept is based on the integration of everyday objects with sensing and networking capabilities to communicate with all devices or services over the Internet [13]. This way, healthcare represents one of the most potential and attractive areas for the implementation of IoT solutions, especially in remote health monitoring, fitness programs, chronic diseases and elderly care. In addition to the possibility of applying big data analytics on the huge amount of data generated by these systems, IoT applications in healthcare are also expected to reduce medical costs, provide specialized treatments and increase the users' quality of life [14]. In order to achieve those goals, many IoT platforms integrated with cloud-based storage solutions, such as Kaa [15], have been developed [16], [17].

Kaa enables to collect data from virtually any sensors and then analyze and visualize it on equipment consoles or mobile devices, and, thereby, delivering end-to-end solutions to the costumers [17].

The Cardiovascular field presents a higher death prevalence, as sudden cardiac arrest and other major cardiac conditions may lead to death. In 2017, Cardiovascular Diseases were the cause of one in three deaths in the United States of America and Coronary Heart Diseases account for the majority of deaths [18]. In Portugal, sudden cardiac arrest represents 20% of deaths in a year, showing a higher death rate than HIV, Breast Cancer and Pulmonary Cancer all summed up [19].

For this matter, it is crucial to be on top of every biosignal that may induce an early diagnosis of major heart conditions that may lead to death. Through the measuring

978-1-5386-9502-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CIC.2018.00061

of the electrical activity of the heart, Electrocardiograms (ECG or EKG) help on the diagnosis of the abnormal heart rhythms, such rhythms may lead to Ventricular Fibrillation and thereafter cardiac arrests [20], [21]. However, early diagnosis of this condition may prevent this outcome.

Focusing on these issues, this paper proposes the creation of an IoT platform for collecting ECG sensors' data with further evaluation of the storage mechanisms. Not only the computational analysis matter but also the visualization of these results may offer a bigger overview of patients' condition.

II. Related Work

The way Big Data is being handled in healthcare is not as efficient as expected. However, some progress is being made towards the efficiency of those systems. As days go by, it becomes clearer the role of big data in healthcare. From personal medical records to clinical trial data, radiology images to bio-metric sensor readings, data is everywhere and must be dealt with great caution. Virtualization and cloud computing are helping the development of new systems for collecting, storing and manipulating these enormous volumes of data [22]. As data keeps increasing, so are the systems and the complexity of the analytic techniques. The mention of unstructured or semi-structured data keeps surfacing in healthcare, the limitations of treating this type of data are diminishing but this field is still challenging [23].

The study in [24] presented a cloud-based system for ECG monitoring and analysis in real-time. The system was built for mobile devices and web browsers and was designed to unravel the time-consumption problems of ECG collected data. As for the monitoring and analysis, the authors implemented algorithms to evaluate and enhance the ECG data, as well as to extract the most relevant features. The proposed system use case was built over a mobile phone sending the ECG data to a web server where the analysis was performed in real-time. The authors also mention the importance of these systems to health in general and the simplicity of their system, which could be used by patients before sending their records for the physicians for assessment and by physicians to help in the diagnosis and the extraction of useful features.

Using Cloud services and an API, the SenseMyHeart application, [25] aimed to tackle the difficulties of mobile monitoring and use of Linux dedicated tools. Focusing on the integration of tools already available, as PhysioToolkit [26] - a large library of software for processing and analysis of physiological signals that is used via command line and requires basic knowledge of Linux-based systems' bash interface - the authors proposed a computing infrastructure, web and mobile based, providing flexibility to researchers and app-developers. The overall goal was to contribute with a "standardized computational backbone" to other projects that intend to study physiological stress or measure Heart Rate or Heart Rate Variability. Although a Windows application was created to be used as an offline analysis system, the Android prototype application was built to work in real-time as a cardiovascular ambulatory monitoring. Connecting directly with the sensor via Bluetooth, the data was then passed to the web service, where it was analyzed and several indicators plotted. The use of a cloud-based service allows the system to present good performance results on the mobile settings. Relying on a stable internet connection, the server-side infrastructure needs to be cautiously created.

In [27] it was attempted to create a system using IoT sensors to predict the number of patients that might suffer a heart attack. Adopting machine learning as the technology to be used for prediction, the author separated the data into three types: fixed, such as sex and age; periodic, like blood sugar or cholesterol level, which is measured by a period of a day or week; and live data, which encompasses blood pressure and heart rate. The first type of data was collected via a web-based platform, while the other two were collected with wearable and IoT devices for home monitoring that can send the data afterward to the remote cloud server. On the other side, for machine learning purposes, the authors used an example dataset from UCI Machine Learning Repository, which was restructured to a format that could be obtained from the previously explained healthcare cloud platform. After applying the k-Nearest Neighbour algorithm, the results seemed to be hopeful. Although the accuracy obtained reached 96%, the number of cases of false negatives was not the best. The author also highlights that, in order to increase the accuracy of the model, the number of false negatives should be reduced.

It is essential to understand that remote monitoring of patients, as well as analyzing patient profile and characteristics makes it possible to identify individuals who would benefit from preventive care or lifestyle changes in order to avoid possible future complications and interventions [2], [3], [28], [29].

All these systems showed promising results but in order to reach even further, the approach proposed here will be able to be adapted to any content in the healthcare environment as well as be able to process the records in near-real time, promoting the early diagnosis of several diseases.

III. Methodology

Figure 1 presents the overall architecture of the proposed platform which will emphasize the comparison between Apache Cassandra and Apache HBase as data storage for medical records. The patient is the data source, from which the ECG is the main data obtained. The system encompasses the data source, the IoT platform, the data storage and an integration of analytic frameworks.

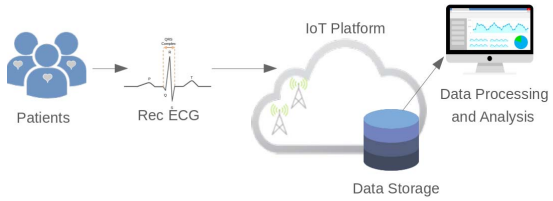

Figure 1. Architecture of proposed system.

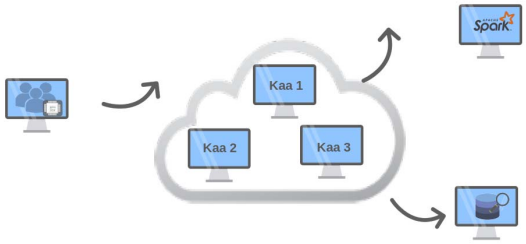

Figure 2. Experimental Scenery of the Platform. Each monitor represents a virtual machine running Ubuntu 16.04 with 2 CPU cores, 4 GB of RAM memory and 20 GB of available space in disk.

A. Apache Cassandra and Apache HBase

Apache Cassandra and Apache HBase are two of the most known and used NoSQL databases. NoSQL databases are commonly used for unstructured data, data generally present in healthcare. These databases are defined by their dynamic schematic, horizontal scalability, manageability, speed and flexibility. These characteristics appraise the use of NoSQL databases when dealing with medical data flowing from sensors [30], [31], [32], [33].

To this end, Apache Cassandra was chosen as one of the databases to be tested along with Apache HBase since the first was already integrated with the IoT middleware and the other has exhibited good performance dealing with big data and is proper for faster streaming mechanisms such as Apache Spark [34]. These two databases are column-oriented, which makes possible to use the time series data in its whole and to query it as needed [31], [33].

B. Kaa IoT and Apache Spark

The proposed platform is based on the Kaa IoT platform deployed in a cluster environment.
Another critical decision to address was related to the storage mechanism. The Kaa platform officially supports Cassandra database as the NoSQL solution for data storage. The connection between Kaa IoT platform and another third party software is established through a driver component designated as log appender. This project allowed to develop an HBase log appender which has already been integrated on the official Github page of the platform.
The implementation of the platform was performed in an experimental environment with virtual machines, in a local area network. Each virtual machine runs Ubuntu 16.04 with 2 CPU cores, 4 GB of RAM memory and 20 GB of available space in disk.
For this experiment setup, the cluster comprises three virtual machines with the specifications referred above (Figure 2). Each virtual machine represents one node of the cluster. In each node, it was deployed an instance of Kaa, Cassandra and HBase. Both Cassandra and HBase clusters were set with a replication factor of 2, which means that for every record it is written two copies. Furthermore, Cassandra consistency was set to ”ONE”, in order to improve the availability of the system.

Another virtual machine was dedicated for data source purposes. All simulated medical sensors were integrated into that machine, which works as an IoT gateway between the sensors and the platform.
Apache Spark and a simple analytic component for fetching data from both databases were developed outside the platform cluster. In order to make them independent from each other, both components were assigned to a dedicated virtual machine.

IV. Results

A. Evaluation Approach

The major contribution of this system is the development of a platform to provide real-time processing mechanisms and its evaluation for data collection and analysis of medical sensors data. One of the most important aspects to evaluate the performance of the proposed system concerns its ability to receive, write and stream the medical records. Another evaluation mechanism was the visualization of the record after being streamed by the platform. The platform needs to be trustworthy and cannot lose any information since medical records can be extremely sensitive and vital.

B. Platform Evaluation

After setting up the experimental environment, it was possible to design the evaluation process of the platform performance. Two standard metrics widely used for that purpose include the throughput and the response time. The throughput represents the number of transactions per second handled by the platform, while the response time is the measure of time between a request operation by a client application and the respective answer by the server. It is important to emphasize the fact that the clocks of the servers are synchronized with the help of the Network Time Protocol (NTP).
For that purpose, four scenarios were designed: the first scenario assessed the limits of the platform; the second scenario tested how the platform reacts with constant write only applications, for different throughput values; the third

scenario concerns the constant write and streaming applications; and for the fourth and last scenario, it was tested constant writing, streaming and analytic applications.
All these scenarios were tested with both databases, Cassandra and HBase, the results are exhibited below.
The first scenario tested the platform with different throughput values and the assessment of these results grounded the results for the next scenarios. In this case, the throughput represents the amount of data sent to the platform per second. The evaluated metrics were based on the response time (RT). The RT presented here is the average response time from several tests in milliseconds.
The second scenario tests were performed based on the results from the first scenario, so, for 10 minutes it was sent to the platform 6000 records per second. The results are shown on Figures 3 and 4 where is exhibited the average response time of the platform for an input of 6000 sensors during an execution time of 10 minutes, per test, for Cassandra (Figure 3) and HBase (Figure 4). The platform presents response time values below 1 second for most samples, in both cases. However, it is possible to observe several higher distinct values, especially with Cassandra. The evaluation was performed with 2000, 6000 and 9000 sensors, despite that, the results here disclosed will only focus on the throughput of 6000 sensors.

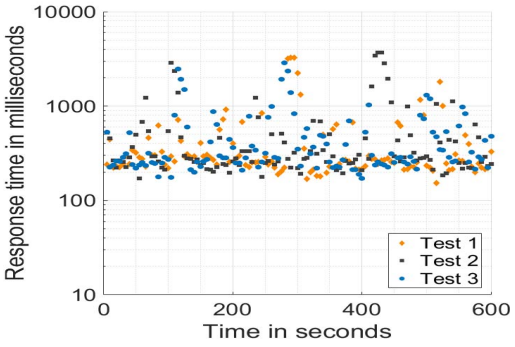

Figure 3. Cassandra

Being guided by the previous results, the third scenario presents only the constant writing and streaming of 6000 records per second (Figures 5 and 6). These figures represent the average response time of the platform for an input of 6000 sensors during an execution time of 10 minutes, per test, for Cassandra (Figure 5) and HBase (Figure 6), along with Spark streaming application. The platform presents response time values below 1 second for most samples. In general, HBase provides lower response time values.

The fourth scenario, where constant writing, streaming and analytic operations were tested, exhibits relevant results. Both tests can be seen on Figures 7 and 8. These figures present the average response time of the platform for an

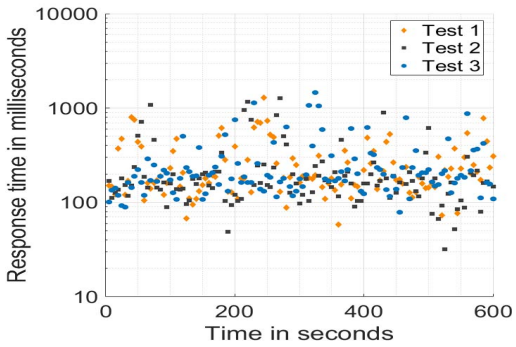

Figure 4. HBase

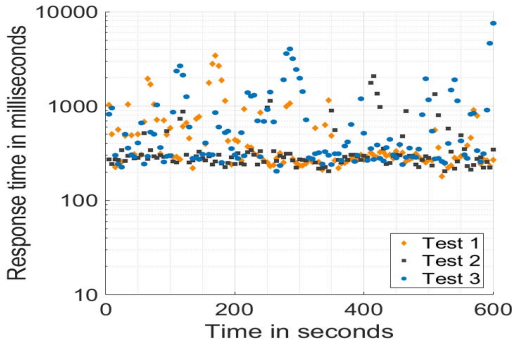

Figure 5. Cassandra

input of 6000 sensors during an execution time of 10 minutes, per test, for Cassandra (Figure 7) and HBase (Figure 8), in a real-world scenario. The platform presents response time values below 1000 milliseconds for most samples. In general, HBase provides lower response time values, while Cassandra presents more peaks in its response time values.

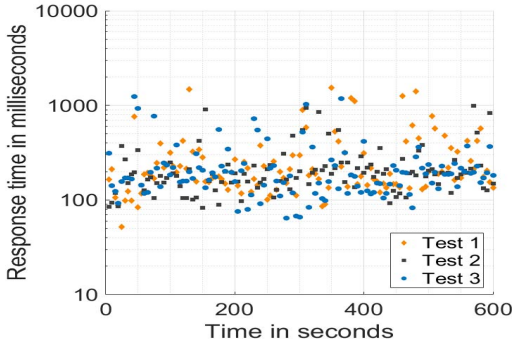

Figure 6. HBase

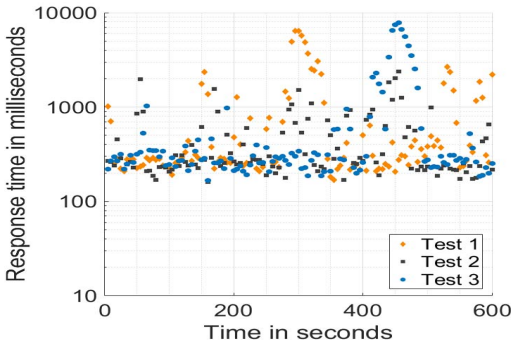

Figure 7. Cassandra

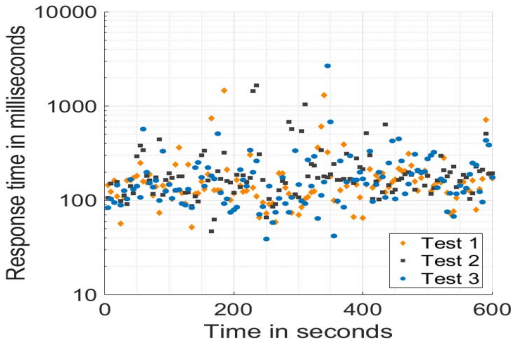

Figure 8. HBase

C. Dashboard Visualisation

The dashboard used for visualizing the records was built on Freeboard.io [35]. Figure 9 exhibits a preview of the data received and streamed by Spark to the dashboard. It is noticed the lack of filtering in the signal which should be done when processing the electrocardiogram. For the purpose of this evaluation, it was used a record from the MIT-BIH Arrhythmia Database [36] which is sampled at 360Hz, meaning it is sent three-hundred and sixty data points per second.

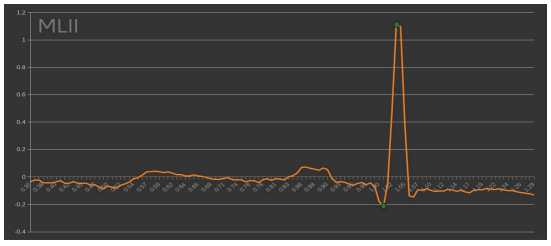

Figure 9. Lead MLII present on the first record from the MIT-BIH Arrhythmia Database [36]

This dashboard has already a built-in functionality that finds the higher and lower peak, which helps to understand where the QRS-complex is, the small green dots in the graph.

V. Discussion

The platform deployed that comprises different technologies exhibited a great performance regarding all the data received. The first scenario allowed to perform an overview of the limits of the platform performance, which made it possible to design further experiments without compromising the system. The second scenario main goal was to explore the performance of the platform for a small, maximum and overwhelming throughput values. The small throughput corresponded to 2000 records per second and presented values lower than 150 milliseconds. The maximum throughput without compromising the platform was around 6000 records per second, as observed in the first scenario, while 9000 records per second represented an overwhelming value, which compromised the platform. Overall, for 6000 records per second, this scenario demonstrates good results despite some values being higher than expected. This condition was observed since the average response time was influenced by some peaks that were due to the internal processes of the Kaa platform and the database such as load balancing, garbage collector, etc. These response time peaks did influence the average response time however it did not influence the amount of stored data.

The three scenarios were fully analyzed and it is possible to observe that the response time increases slightly when the platform needs to write, stream and use analytic tools.

The results showed that the platform with HBase was lower response time values and, consequently, less loaded queues. HBase presented in all the three scenarios fewer peaks than Cassandra.

The impact of the fourth scenario on the platform performance was slightly reflected in the increased response time of the platform with Cassandra. However, the impact on the platform with HBase was not noticeable.

This system can bring a different perspective regarding the overall subject as far as it encompasses all the open-source tools available and creates a platform which can get above average results. Although full integration of both systems has yet to be accomplished, these results can prove that a fully functional platform can be made for healthcare purposes and to increase the performance when managing and analyzing sensing data.

When observing the dashboard, it can be seen the lack of preprocessing of the signal, which may induce errors when dealing with such sensitive data. It is, therefore, crucial to creating a preprocessing stage before the analysis and visualization of this medical data. Although it is known that several medical data may not need a preprocessing stage, sensoring data do need that additional stage.

VI. Conclusion

The main goal of this paper was the development of an IoT platform designed for data collection and analytic purposes for medical sensors, in the healthcare environment, more specifically electrocardiograms. A system capable of handling big data provides an excellent opportunity for big data analytics to get in action, in order to originate useful insights based on all that data. In the healthcare sector, this means that not only it will be possible to provide high quality and personalized care to patients, but it also enables the development of preventive care systems. Furthermore, the application of IoT technologies allows establishing a connection between all devices and medical sensors in the respective healthcare environments, which is especially useful for remote patient monitoring. The development of a platform capable of integrating both the IoT and Big Data concepts was the main motivation of this paper.

The results obtained proved that the platform is suitable for the deployment in a healthcare environment, where a large amount of data is generated in small time intervals. The implementation of HBase in Kaa platform represents a good contribution to the system, because not only presents an alternative with slightly better response time values than Cassandra but also provides a great opportunity for analytic applications.

Future scenarios should be tested since all of the components are horizontally scalable and it is safe to say that the overall performance of the platform would increase with the addition of more nodes. Nevertheless, the assessment of the platform with more nodes would be an interesting subject of study.

On the other side, more analytic tools are under development likewise, a real-time deep learning classification of electrocardiograms is already being developed to be implemented on the platform built.

Knowing how fast medical data is growing it starts to be a necessity more than a need to implement scalable and reliable systems which can disclose all the problems large and fast volume of data may cause.

References

[1] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The 'big data'revolution in healthcare," *McKinsey Quarterly*, vol. 2, p. 3, 2013.

[2] D. Saidulu and R. Sasikala, "Understanding the challenges and opportunities with big data applications over a smart healthcare system," *International Journal of Computer Applications*, vol. 160, no. 8, 2017.

[3] 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.

[4] S. Salas-Vega, A. Haimann, and E. Mossialos, "Big data and health care: challenges and opportunities for coordinated policy development in the eu," *Health Systems & Reform*, vol. 1, no. 4, pp. 285–300, 2015.

[5] S. Patel and A. Patel, "Abig data revolution in health care sector: Opportunities, challenges and technological advancements," *International Journal of Information*, vol. 6, no. 1/2, 2016.

[6] K. Jee and G.-H. Kim, "Potentiality of big data in the medical sector: focus on how to reshape the healthcare system," *Healthcare informatics research*, vol. 19, no. 2, pp. 79–85, 2013.

[7] "Apache Hadoop documentation," http://hadoop.apache.org/docs/stable/, Accessed: 2018-02-01.

[8] "Apache HBase reference guide," http://hbase.apache.org/book.html, Accessed: 2018-02-01.

[9] "Apache Cassandra documentation," http://cassandra.apache.org/doc/latest/, Accessed: 2018-02-01.

[10] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," *Computer science review*, vol. 17, pp. 70–81, 2015.

[11] "Apache Spark documentation," https://spark.apache.org/docs/latest/, Accessed: 2018-02-01.

[12] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen *et al.*, "Mllib: Machine learning in apache spark," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[13] A. Whitmore, A. Agarwal, and L. Da Xu, "The internet of things—a survey of topics and trends," *Information Systems Frontiers*, vol. 17, no. 2, pp. 261–274, 2015.

[14] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[15] "Kaa IoT platform documentation," https://kaaproject.github.io/kaa/docs/v0.10.0/, Accessed: 2018-02-01.

[16] A. M. Nagib and H. S. Hamza, "SIGHTED: A framework for semantic integration of heterogeneous sensor data on the internet of things," *Procedia Computer Science*, vol. 83, pp. 529–536, 2016.

[17] R. Suvarna, S. Kawatkar, and D. Jagli, "Internet of medical things [iomt]," *International Journal*, vol. 4, no. 6, 2016.

[18] B. EJ, "Heart Disease and Stroke Statistics - 2017 Update," *Circulation*, vol. 135, pp. 2017–2018, 2017.

[19] D. J. L. Gomes, "A Morte Súbita Cardíaca - Fundação Portuguesa Cardiologia." [Online]. Available: http://www.fpcardiologia.pt/a-morte-subita-cardiaca/

[20] G. J. Tortora and B. Derrickson, *Principles of Anatomy & Physiology 14th Edition*, 2014.

[21] A. C. Guyton and J. E. Hall, *TextBook in Medical Physiology*. Elsevier Saunders, 2006.

[22] B. Feldman and E. M. Martin, "Big Data in Healthcare Hype and Hope," no. October, 2012.

[23] 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.

[24] H. Xia, I. Asif, and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis," *Computer Methods and Programs in Biomedicine*, vol. 110, no. 3, pp. 253–259, 2013.

[25] P. M. Pinto Silva and J. P. Silva Cunha, "SenseMyHeart: A cloud service and API for wearable heart monitors," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-November, pp. 4986–4989, 2015.

[26] "The WFDB Software Package." [Online]. Available: https://physionet.org/physiotools/wfdb.shtml

[27] F. Ahmed, "An Internet of Things (IoT) Application for Predicting the Quantity of Future Heart Attack Patients," vol. 164, no. 6, pp. 36–40, 2017.

[28] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[29] S. I. Lee, H. Ghasemzadeh, B. Mortazavi, M. Lan, N. Alshurafa, M. Ong, and M. Sarrafzadeh, "Remote patient monitoring: what impact can data analytics have on cost?" in *Proceedings of the 4th Conference on Wireless Health*. ACM, 2013, p. 4.

[30] N. Bhojwani and A. P. V. Shah, "A survey on hadoop hbase system," *Development*, vol. 3, no. 1, 2016.

[31] L. George, *HBase: the definitive guide: random access to your planet-size data.* " O'Reilly Media, Inc.", 2011.

[32] H. Wang, J. Li, H. Zhang, and Y. Zhou, "Benchmarking replication and consistency strategies in cloud serving databases: Hbase and cassandra," in *Workshop on Big Data Benchmarks, Performance Optimization, and Emerging Hardware*. Springer, 2014, pp. 71–82.

[33] M. Brown, *Learning Apache Cassandra*. Packt Publishing Ltd, 2015.

[34] R. C. Taylor, "An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics," in *BMC bioinformatics*, vol. 11. BioMed Central, 2010, p. S1.

[35] "Freeboard - Free dashboard for Internet of Things visualization," http://freeboard.io/, Accessed: 2018-08-02.

[36] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database." *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society*, vol. 20, no. 3, pp. 45–50, 2001.