Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

IoT, Cloud Computing and Big Data Techniques for Smarter Healthcare

Nada Chendeb Taher*[a], ImaneMallat[a],Nazim Agoulmine[b], Nour Mawass[c]
[a]Lebanese University, Faculty of Engineering, Tripoli, Lebanon
[b]COSMO, IBISC Laboratory, University of Evry, France
[c]Normandie University, UNIROUEN, LITIS

Abstract: With the increasing number of connected things to the internet (IoT), the volume of the generated data and the rate at which it is generated are more and more increasing.In general, most of the generated data from these IoT devices is stored on some cloud infrastructure to insure scalability and continuous ease of access to it.With this situation, the concept of Big Data recently appears to lead the Artificial Intelligence and the Data Mining domains. In fact, if we use this 'Big Data' correctly, we could turn it into 'Big Value'. Actually, the essential target of the emerging Big Data technologies is to provide techniques and tools to store large amount of complex data and prepare it to be analyzed and processed in order to get insights and predictions that could offer new opportunities towards better future.In this context, we have to deal with two main issues: the real-time analysis issue introduced by the increasing rate at which data is generated from IoT devices, and the long-term analysis issue introduced by the accumulation over time of huge volumes of data.

In smart healthcare applications, these two issues appear clearly. In fact, medical sensors collect health related data from patients and send it to the cloud. This data should be analyzed permanently in real time to take appropriate decisions and act accordingly to save the patient's life for example. The same data accumulated over time from different patients constitutes the bank of training data to build accurate machine learning model in order to perform smarter future disease prediction.

In this paper, we propose an IoT-Cloud based solution for real-time and batch processing of Big Data. We use a Raspberry pi to replace the IoT device and generate data in real time. The objective is to provide a smart healthcare application into the cloud integrating IoT and Big Data techniques provided by the cloud operator to perform smart ECG analysis and early detection of any ECG anomaly. The solution was implemented and tested in AWS Amazon Cloud, it worked well and results show that the processing performance in term of response time for both long-term and real-time analysis is always guaranteed once the cloud resources are well provisioned.

Keywords: *Cloud Computing; IoT; Big Data Analytics; Healthcare; real-time analysis, batch processing*

I. INTRODUCTION

From the down of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days and the pace is accelerating [1]. Today, it is hard to imagine any activity that does not generate data. Also, with the era of IoT, we are increasingly surrounded by sensors that collect and share data in many devices and products. With the datafication of everything, comes Big Data.

Cloud Computing, IoT and Big Data are distinct disciplines that have evolved separately over time. However, they are increasingly becoming interdependent. The concept of Big Data and IoT has been around for many years, but its mainstream application started only recently [2]. The concept of cloud computing also traces back to the 1960s 'intergalactic computer network' and has since then evolved and passed through many stages to become today a mainstream commercial necessity [2].

Demand for Big Data is calling for the adoption of both IoT and Cloud platforms. With IoT, the amount of data will obviously dramatically increase. The adoption of IoT and Big Data compels a move towards Cloud technology. Therefore, if we need to transform the IoT data and utilize its potential, we need first to fully embrace Cloud-based systems.

IoT, Big Data and Cloud Computing form together a coherent ring; IoT embraces the cloud and in parallel generates Big Data that integrates again the cloud computing services. These three technologies are strongly interconnected as shown in Figure 1.

The convergence of IoT, Big Data techniques and Cloud Computing can provide new opportunities and applications in many sectors. Today, many companies integrate IoT and Cloud technologies, in their way to digitize and analyze data, and finally get predictions and insights, that promote businesses. We can find many applications that deal with these technologies in many domains such as healthcare, industry, business, marketing and many others domains.

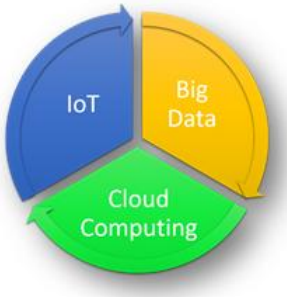

Fig. 1 IoT, Big Data Analysis and Cloud Computing interconnection

For example, in the healthcare domain, the humongous amount of personal data collected through personal records and sensors from different medical devices needs a platform for storage. This is where Cloud Computing becomes important. Further analysis of these data can enable smarter healthcare system with personalized diagnostics and assistance.

In this context, any application dealt with the IoT-Cloud technology can take one or both of two ways: the (1) real-time analysis that helps to find out irregularities in the collected data and act as fast as possible to prevent any undesired scenario, or the (2) long-term analysis that uses the massive

data collected from IoT devices and utilizes the insights to identify the future trends and opportunities.

With the real-time analysis, we face the "Velocity" challenge. In major applications, we have to process data and get information and decision before the next data is generated. Moreover, in critical applications such in the medical domain, we have to act as fast as possible to save the patient's life. While with the long-term analysis that requires batch processing, we face the "Volume" challenge; with the increasing volume of data, where to store the huge amount of data and how to process it accurately and in a reasonable amount of time?

The two different ways of processing data lead to the necessity of an IoT-Cloud based solution for Big Data analytics that deals with both the volume and the velocity challenges as the diagram of Figure 2 shows.

Some Cloud operators offer all the tools to build our model, but which one is the best one? What is the best performance/cost tradeoff? In this work, we aim to build such a model into a cloud operator and answer the above questions.

A typical application of the proposed solution/model is in the healthcare domain where the ECG sensors measure heart parameters and send them to the cloud for real-time analysis in order to take critical decisions or notify doctors/health specialists. These decisions/notifications are simply the result of batch processing/machine learning applied on the data accumulated over time. Implementing such as solution over the proposed model seems also very interesting and beneficial in the medical field.

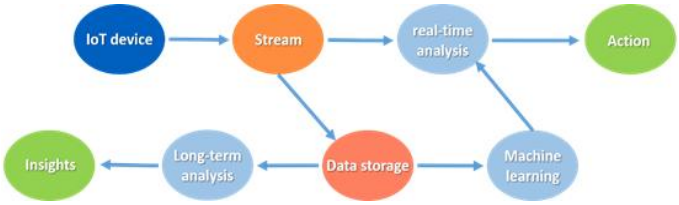

Fig. 2: A diagram for an IoT-Cloud-Big Data model for real-time and long-term analysis

In this article, and after some preliminaries and literature review in section 2, we will introduce our solution for an IoT-Cloud based model for Big Data analytics with the services to use from AWS Cloud operator in section 3. After the implementation, we apply the solution to a healthcare scenario in section 4 and finally we conclude in section 5.

II. PRELIMINARIES AND RELATED WORKS

1) What is Big Data?: There is no stable definition for Big Data. We can use Big Data to describe a massive volume of both structured and unstructured data that is so large, increasing very fast and that is difficult to process using traditional tools, databases and software techniques. To deal with this kind of data, a multitude of tools and frameworks become recently available including the famous Hadoop ecosystem. Basic techniques behind these tools and frameworks are distributed, parallel and cluster computing.

Hadoop is an open-source framework that allows to store and process large datasets in parallel and distributed fashion. It runs on clusters of commodity servers and can scale up to support thousands of hardware nodes and massive amounts of data. Consequently, Hadoop became a data management platform for big data analytics.

From this ecosystem, we will use the Hadoop 'HDFS' as Hadoop Distributed File System, that is a parallel and distributed storage for storing our data into the cloud and the 'Spark' platform, an open-source big data processing framework built around speed, ease of use, and sophistical analytics. We select Spark because it supports SQL queries, streaming data, machine learning and graph processing; it can therefore ensure a real-time processing and sophistical analytics like 'Machine Learning' and 'data mining'. Spark also offers a shell for python called 'PySpark'. 'PySpark' includes many libraries like 'MLlib' library for 'Machine Learning' algorithms, and 'pandas' library for data analysis and exploration.

2) Cloud Computing: Cloud Computing is a paradigm in which any user with internet connection can rent computing resources as needed from a cloud operator owning large datacenters and offering services. This is mainly an economic revolution in the IT/Networking field based mainly on the virtualization technology. Cloud computing services cover a vast range of options now, from the basics of storage, networking, and processing power through to natural language processing and artificial intelligence. Three of the main benefits of cloud computing are self-service provisioning, elasticity and Pay per use.

3) Related works: With the exponential growth of data pertaining to the health and speedily generated from IoT and sensing devices, healthcare industries are suffering from storing, processing and providing insights from such large amount of data. This health data explosion makes Big Data solution a necessity in healthcare area, nowadays. Big Data solutions contribute to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general.

In the last decade, researches quickly evolved in the domain of Big Data and healthcare, given its importance in medical improvement. Several works care about Cloud-based infrastructure for Big Data long-term analytics, while several other works care about real-time or data stream processing. In the following, we will discuss some of these researches.

As an attempt to enable more accurate estimation of storage and memory for the real-time online health analytics and retrospective analytics components, IBM Research and Development Center in cooperation with the University of Ontario Institute of Technology introduced since 2014 a project called 'Artemis Project' in the area of the neonatal intensive care units (NICU). A pilot project was deployed at the SickKids hospital in Toronto. Artemis Cloud is capable of gathering physiological data from a vast variety of medical devices and monitors in a secure way and performing the online analytics within the cloud [3]. In addition to real-time analytics, Artemis Cloud is able to provide at-rest analytics for stored data with the capabilities of visualization both the real-time and historical data. Artemis Cloud platform was modelled as M/M/m/m queuing system with no extra capacity than service facilities. Based on this performance model, important performance metrics such as mean number of patient in NICU, mean patient residence time, mean number of required medical algorithms and blocking probability were characterized and discussed, hence leading to identify the amount of required storage, memory and computation power for analytics and real time components respectively.

Recently, Big Data stream computing has been studied in order to improve the quality of healthcare services and reduce costs by capability support prediction, thus making decisions

in real-time [4]. Van-Dai Ta, Chuan-Ming Liu, and Goodwill Wandile Nkabinde, introduced a generic architecture for big data healthcare analytic by using open sources, including Hadoop, Apache Storm, Kafka and NoSQL Cassandra. Using Kafka, Storm, and NoSQL Cassandra for stream computing, to combine with existing Hadoop-HBase framework, this proposed architecture can support for healthcare analytics by providing batch and stream computing, extendable storage solution and effective query management [4].

Another proposed architecture for big data in healthcare was presented by Megan Sheeran and Robert Steele [5]. This framework includes four layers. The Health Data Sources layer describes the various sources where the health-related data originates from. The next layer uses Big Data technologies to upload these datasets into software with the ability to store large amounts of data, as a step towards carrying out analytics on these data. The third layer is the Big Data Analytics Layer, which includes the mechanisms for performing analytics on the data; the data warehousing, data mining, and machine learning is done at this layer. The fourth and final layer describes the implementation of the end goals of analytics on Health Big Data.

As an application to Big Data potential in the healthcare domain, Prashant Johri, Tanya Singh, Sanjoy Das and Shipra Anand [6] proposed a Big Data architecture based on Hadoop platform, Hive to solve SQL queries and transform unstructured data to structured format, and the R programming language that generates statistical result and graphics. They used the datasets of two different countries, India and China, affected from diabetes and then visualized their systolic blood pressure to gain an insight of the coming years; if the current trend of high blood pressure continues to grow at an alarming rate, it would be much more difficult to cure diabetic patients as well as preventing the next generation from diabetes. Using predictive analytics, they were able to find the future condition of both the countries and to improve it they will try to reduce the factors which cause it [6].

Based on these researches and studies, we can deduce the importance of Big Data in the medical field. Integrating IoT and Big Data solutions means smarter healthcare services. This integration needs a solution that cannot be achieved without cloud infrastructure. Therefore, in the following, we will present our Cloud-based solution that deals with both real-time and long-term processing of medical data. Real-time analysis receives IoT generated data and predicts anomalies, while long-term processing helps to analyze and predict the trends of diseases in order to reduce its risks. We note also that this solution could be applied to any application having these two aspects.

III. IOT-CLOUD SOLUTION FOR BIG DATA

Nowadays, streaming data is seen and used everywhere. IoT devices constitute one main source of streaming data among others. As the speed and the volume of this type of data increases, the need to perform the data analysis in real time with machine learning algorithms and extract a deeper understanding from the data becomes ever more important. In parallel, to satisfy the increasing velocity and volume even the complexity of generated data, the use of Big Data tools and techniques that ensure parallelism in computing, scalability and reliability becomes a necessity.

In this context, batch processing is needed to analyze and explore a large amount of complex and incomprehensible data and finally get insights from this data, on the other hand the real-time processing is needed to stream new data, predict anomalies and finally take the appropriate decisions. In our application, we might want a continuous monitoring system that receives data from a huge number of ECG devices, to detect sudden anomalies, so that we can react to the anomaly in real-time to save the patient. This system need to be built around a machine learning algorithm to predict any abnormality from The ECG coming data in real time. Here lies the importance that this system also deals with the batch processing to first analyze data for the machine learning model, and to finally test the accuracy of the model.

Consequently, and based on the importance given for both the real time and the long-term analysis in the majority of the domains today, we created an IoT-cloud based system using the 'Amazon' Cloud that provides all Big Data techniques and tools we need to perform such a system, and at an affordable cost without the necessity to procure hardware or to maintain infrastructure. To perform our solution on Amazon we referred to a set of Amazon Web Services (AWS) that can be connected between them. Some of these services are for capturing data streams from IoT devices, others for compute and processing and some others for storage and notification. In the following we will describe the services used in our model.

A. AWS services used in our model

1) AWS IoT core: Amazon IoT service is used to connect IoT devices, receive data from these devices using 'MQTT' protocol and publish the messages to a specific 'topic'. Additionally, AWS IoT 'rules' applied on the received data, gives IoT-enabled devices the ability to interact with other services. Rules are analyzed, and actions are performed based on the MQTT topic stream.

*2) Amazon Kinesis Stream:*Amazon Kinesis Streams can continuously capture and store terabytes of data per hour and hundreds and thousands of sources. With amazon kinesis streaming data can be ingested, buffered and processed in real-time, so insights can be derived in seconds or minutes instead of hours or days.'Shard' is the base throughput unit of an Amazon Kinesis stream. An Amazon Kinesis stream is made up of one or more shards. Each shard provides a capacity of 1MB/sec data input and 2MB/sec data output. Each shard can support up to 1000 write and 5 read transactions per second. The number of shards needed within the stream is specified based on the throughput requirements.

*3) Amazon Elastic Compute Cloud (EC2):*Amazon EC2 provides scalable computing capacity. Amazon EC2 can be used to launch as many or as few virtual servers as we need. We only pay for EC2 instances we use per hour depending on which EC2 instance type is used.

4) Elastic MapReduce (EMR): Amazon EMR is a highly distributed computing framework to easily process and store Big Data quickly in a cost-effective manner. Amazon EMR uses 'Apache Hadoop', to distribute the data storage and processing across a resizable cluster of Amazon EC2 instances and allows to use the most common 'Hadoop' tools such as 'Hive', 'Pig', 'Spark' and so on. With Amazon EMR, we can add core nodes at any time to increase the processing power.

5) Simple Storage Service (S3): Amazon S3 provides developers and IT teams with secure, durable, highly-scalable cloud storage.

6) Short Notification Service (SNS): SNS is a fully managed push notification service that allows sending individual

messages to large numbers of recipients.

B. Our proposed model

These services already described should be carefully connected to perform our IoT-Cloud based solution for real-time and batch processing of Big Data as presented in Figure 3. First, we have to create a Spark cluster of specific EC2 instance type using Amazon EMR. After uploading data to an S3 bucket, data is pulled from Spark cluster to train a machine learning network. A Raspberry Pi (Rpi) that plays the role of an IoT device is connected to the Amazon IoT core via MQTT protocol. Data sent from the Rpi is published to a specific topic where a kinesis rule is applied to push data into the kinesis stream shard. In the EMR Spark cluster, streaming data is pulled to be predicted with the machine learning model already built. In case of any anomaly, a notification is sent with Amazon SNS to a specific mobile number.

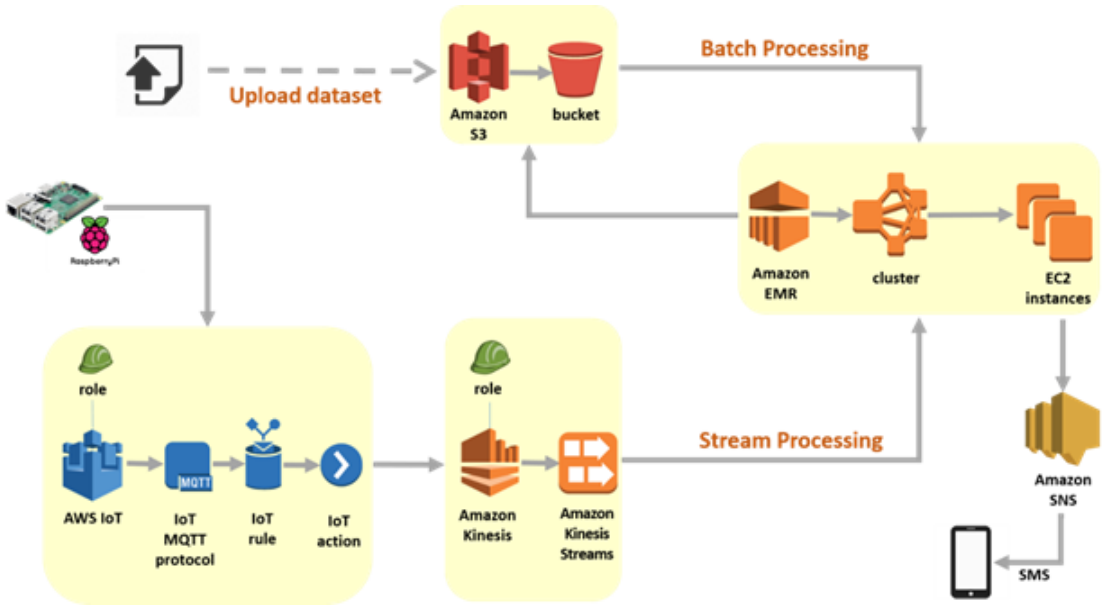

Fig. 3: Our IoT-Cloud solution for Big Data analytics

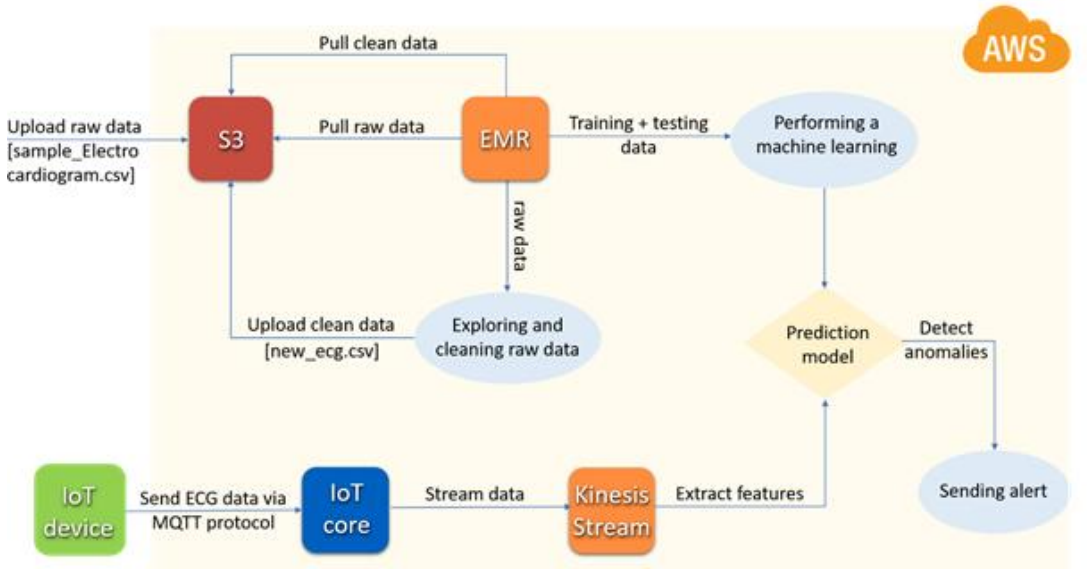

Fig. 4: IoT-Cloud based solution applied on ECG data

C. Test phase for performance analysis

To test our solution, we used a typical big dataset; it was selected to validate our work and to do the performance/cost analysis of the complete solution in the cloud. This is the dataset used for The Third International Knowledge Discovery and Data Mining Tools Competition [7].

To test our model, we used the full data file for training network, the data file with corrected labels for testing the machine learning network and the unlabeled data file to stream data from the Rpi thus simulating new data that should be analyzed in real-time.

We used the 'logistic regression', a supervised machine learning model, on this dataset to distinguish between normal and abnormal data. The Machine learning model was inspired from a tutorial applied on the same dataset testing different algorithms in Apache Spark [8].

For the performance analysis compared to the cost of the solution, we tried different EC2 instance types and every time a different number of instances by cluster, the objective being to choose the cheapest and the most performant one (we need a

compromise between the processing time and the cost).

We noticed also that some configurations are insufficient for big scale of the data and cannot achieve the processing requirements. Once the resources are scaled up, the processing time becomes acceptable.

Concerning the response time of the real-time analysis, in the data we use, one record doesn't exceed 150 Bytes, so it was enough to use only one Shard, on condition that the period between two sent records will be in minimum 1 ms, while the time between two received records will be in minimum 0.2 sec. The response time was very low and do not exceed the ms. If we need to send or receive bigger records, or faster, we have to increase the number of Shards and the performance is always guaranteed. In this way, Kinesis Stream ensures the scalability to the model and the stream processing remains in real-time but we pay more.

IV. HEALTHCARE APPLICATION ON ECG DATASET

After building and testing our solution, and after analyzing the performance we may obtain in the cloud for Big Data analytics, we can confirm that any level of the required performance could be obtained with the infinite resources available in the cloud, this is mainly a question of cost and good provisioning of resources. Now, it is time to focus on our application and apply it to a medical dataset containing information about ECG signal features captured from a Korean population, 'ECG-VIEW' dataset [9]. We note that in order to get good insights from Big Data, this data should be real, accumulated over time, with well-defined hypothesis drawn by specialists in the domain.

The dataset contains mainly seven tables, the 'sample-Electrocardiogram' data table contains the extracted ECG parameters and consist of 13 columns: "personid" that represents the patient id, the test date "ecgdate", the clinical department "ecgdept", "ecgsource" that indicates the source of the ECG record, RR interval, PR interval, QRS duration, QT interval, QTc interval calculated using the QT and RR intervals, "P-wave-axis" that is the degree of P wave axis, "QRS-axis" that is the degree of QRS axis, and "T-wave-axis" that represents the degree of T wave axis. A portion of this data table is shown below.

This dataset, 'ECG-VIEW', offers the opportunity to study the electrocardiographic changes based on the ECG features: RR interval, PR interval, QRS interval, QT interval, QTc interval, P wave access degree, QRS axis degree, T wave axis degree, and then detect any ECG anomaly.

Based on the evaluation of each ECG parameter, we can think about a machine learning model that receives the ECG features and distinguishes between the normal or abnormal ECG. If we will apply the supervised learning, 'logistic regression', to deal with the model already used and implemented in the previous section, we will first clean and preprocess data in order to add 'Targets' to the existing dataset. Then the dataset will be divided into training and testing data; 30 percent of all data for testing and 70 percent for training the 'logistic regression' network.

The block diagram shown in Figure 4 presents in details how to apply the cloud based solution for real-time and batch processing on the 'ECG-VIEW' dataset.

What is mainly different from the previous diagram is the data cleaning and pre-processing phase. First we create a bucket on the amazon S3 in where we upload the 'sample-Electrocardiogram' data. Then we create a Spark based cluster with 3 EC2 instances on EMR. It is enough with this type/volume of data to use the c5.xlarge EC2 instance type that consists of 4 CPU and a RAM of 8 GB. After we create the EMR cluster we pull the 'sample-Electrocardiogram' data to be cleaned and pre-processed before performing the machine learning model.

Once the raw data is cleaned and transformed into the 'new-ecg.csv' dataset, we can then perform the machine learning model that takes 70 percent of the 'new-ecg.csv' records for training and the rest for testing the model built.

After we build the logistic regression model, and test its accuracy we note that the accuracy is near 73 percent. This accuracy is less than required and this means that either the training data is not enough to well train the network, either the machine learning model doesn't deal with the data. We think in the case of this type of data which is numeric, the logistic regression model could be applied without problem but the volume of the training data was not sufficient, so to increase the accuracy, we have to increase the volume of training data.

As a summary, we first upload the ECG raw data to an S3 bucket. We create an EMR Spark cluster with 3 c5.xlarge EC2 instances. We pull the ECG raw data into the Spark cluster to be prepared and cleaned, and then the modified data is stored again into the S3 bucket. To perform the logistic regression model, we have to pull the new modified data that is divided into training and testing data. After training the network, we test the accuracy of this model. And now the model becomes ready to predict new ECG records. We create a 'test-stream' kinesis stream with one shard and in IoT core we create a rule that push all published data on the 'kinesis-topic' topic to the 'test-stream' shard. On the spark cluster we read The ECG data continuously from the 'test-stream' shard. Once an ECG record is read, it is predicted using the logistic regression model and a notification is sent to a specific phone number in case of abnormality.

Our cloud based model using the three technologies IoT, cloud computing and Big Data was applied to a dataset containing ECG data. It provides a scalable solution for real-time ECG abnormality detection, even for the long-term analysis, like studying the ECG changes caused by medications and diseases. The data volume we used was not enough to have a sufficient accuracy and this can be corrected if we have more records.

CONCLUSIONS

In conclusion, the IoT cloud based model for Big Data we built can reach the desired goal in terms of response time and accuracy, with a low cost relatively. We always have a cost/performance tradeoff; the increase in complexity, speed, and volume of data leads to using more Cloud resources with higher features and hence paying more. So we have to define the exact needs on the Cloud to reduce costs and then make an efficient model.

Our model can be a baseline for bigger and more complex applications. The cloud-IoT built model can be also applied to a more general medical application including more IoT devices that collect many types of signals (ECG, EMR, blood pressure, temperature, etc..). These sensors can be connected to one or more gateway. Here we have to compute again the resources we need on the amazon to deal with this case.

We think also that to make value from this achieved work

and turn this model into a utility for hospitals and healthcare specialists, some meetings are needed to discuss the needs and build cooperatively real applications. In fact, such work could not succeed without two parties: 1) the specialists that have the full knowledge of the medical data to be analyzed with the objectives of the data analytics to be done on this data and 2) the engineers/researchers that develop the convenient code and algorithms to achieve the right machine learning model and stream analysis for prediction and provide dynamically the needed Cloud resources according to the volume/velocity/scale of the application.

References

[1] Eric Schmidt

[2] Sampriti, Sarkar (2017) Convergence of Big Data, IoT and Cloud Computing for Better Future. Analytics Insight.

[3] HamzehKhazaei, Carolin McGregor, Mikael Eklund, Khalil El-Khatib, Anirudh Thommandram2014Toward a Big Data Healthcare Analytics System: A Mathematical Modeling Perspective. 2014 IEEE World Congress on Services, pp. 208-215.

[4] Van Dai Ta, Chuan-Ming Liu, Goodwill Wandile Nkabinde (2016) Big data stream computing in healthcare real-time analytics. 2016 IEEE International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp. 37-42.

[5] Sheeran, Megan and Steele, Robert (2017) A framework for big data technology in health and healthcare. 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCOM), pp. 401-407.

[6] Prashant Johri, Tanya Singh, Sanjoy Das, Shipra Anand (2017) Vitality of big data analytics in healthcare department. 2017 International Conference on Infocom Technologies and Unmanned system (Trends and Future Directions) (ICTUS), pp. 669-673.

[7] Knowledge Discovery and Data Mining Tools Competition 1999 Data [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99

[8] Spark Python Notebooks [Online]. Available: https://github.com/jadianes/spark-pynotebooks/lob/master/README.md

[9] ECG-View [Online]. Available: http://www.ecgview.org/

[10] Young-Gun Kim, Dahye Shin, Man Young Park, Sukhoon Lee, Min Seok Jeon, Dukyong Yoon, Rae Woong Park (2017) ECG-ViEW II, a freely accessible electrocardiogram database. PLOS ONE.

Acknowledgement

This research was supported by The Lebanese University and CNRS Lebanon. Part of this work was also conducted in the frame of the PHC CEDRE Project N37319SK.