Map-Reduce Based Algorithm for the Analysis of Vital Variables of Neonatal Patients

Sipan Arevshatyan
Department of Computing Engineering, Universitat Politècnica de València
sipan443@gmail.com

Houcine Hassan
Department of Computing Engineering, Universitat Politècnica de València
husein@disca.upv.es

Carlos Domínguez
Department of Computing Engineering, Universitat Politècnica de València
carlosd@disca.upv.es

ABSTRACT

This paper describes neonatal healthcare problems and shows ways to analyse the information using modern technologies like Big Data, Hadoop, and Map Reduce programming in order to help doctors to solve these problems. In the work we have evaluated 4 characteristics of 10 patients by designing an algorithm to analyse the data. The paper consists of four parts: Introduction, related work, application problem and evaluation and conclusion. In these chapters, we explain the concept, characteristics and need of Big Data, state the problem in neonatal health care, provide useful information about related works, show that Big Data do exist in hospitals and conclude the work.

Keywords

Neonatal health care, Big Data, Data analysis, Hadoop, Map Reduce programming

1. INTRODUCTION

Due to the human activities, a lot of information is being created every day. One of the most interesting field of these activities can be considers health care and particularly neonatal health care as there are a lot of neonates who are born with some health problems.

Thus, doctors need to use the information coming from NICU tools in order to make right decisions. Sometimes the information is so big and it is useful to use modern technologies in order to achieve good results. In this point of view, we can say that doctors usually are dealing with Big Data. Big Data is a term for massive data sets having large and complex structure with the difficulties of capturing, storing, searching, sharing, analysing, transferring and visualizing for the future processes [11]. Big data in neonatal health care can be described by three main components: variety, velocity and volume.

Volume: The quantity of generated data in NICU is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered Big Data or not. The name ‘Big Data’ itself contains a term related to size, and hence the characteristic.

Velocity: In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development. For example, smart infusion pumps (SIPs) contribute can provide more than 60 different types of data every 10 seconds.

Variety: The type and nature of the data. This helps people who are associated with and analyse the data to effectively use the data to their advantage and thus uphold its importance.

2. RELATED WORK

The market for healthcare services has increased exponentially. This is due to the growing tendency for personal healthcare to move away from the traditional hubs of healthcare, such as hospitals and clinics, to the private home and especially the mobile environment. In most developed countries an aging population contributes to the growth in the demand for distributed healthcare services. As a result of the nature of healthcare, the precision and real-time delivery of data is crucial. To fulfil all these requirements, advanced and smart technologies should be applied.

One of the very smart technologies was designed by Carolyn McGregor, University of Ontario Institute of Technology, Canada. The platform is called Artemis. Artemis is an online health analytics platform that enables concurrent diagnoses of multiple patients through real-time analysis of multiple data streams [1][7]. It supports acquisition and storage of patients´ information for the purpose of online analytics. Artemis is currently implemented in and used to help sick children in Ontario, Canada and the research team is going to deploy the platform in other cities of Canada as well as in China and Australia.

Another research made in USA by Rollins School of Public Health, Emory University, USA shows the importance of maternal health during pregnancy [2]. The research was held in 13 states of the USA. It states that smoking increased infant length of stay by 1.1%. NICU infants cost $2496 per night while in the NICU and $1796 while in a regular nursery compared to only $748 for non-NICU infants. Multivariate analysis is used to estimate the relationship of smoking to probability of admission to an NICU and, separately, the length of stay for those admitted or not admitted to an NICU.

*ICAIT'*16, Oct. 6–8, 2016, Aizu-Wakamatsu, Japan.

3. APPLICATION PROBLEM AND EVALUATION

3.1 Health Care and Big Data

Probably some years ago, one might not expect these two areas might to even be mentioned in the same sentence. But now they are coming together and tend to change the face of the medicine. The coming together of healthcare and Big Data means higher tech medical solutions for the general problems in medicine. Particularly, high tech medical solutions could be implemented in neonatal health care and medicine [5].

In this article the problem is related to the analysis of vital variables of neonates in hospital.

A neonatal intensive-care unit (NICU), also known as an intensive care nursery (ICN), is an intensive-care unit specializing in the care of ill or premature new born infants. While a child is inside of NICU medical devices generate a lot of information. Devices monitor physiological data streams that reflect the functioning of vital organs while others provide ventilation support. Sometimes the information is so big that we can say that doctors deal with Big Data. In this point of view it is important to gather the information and use it in order to make decisions:

The information coming from NICU can vary. Here are some examples of the Bid Data in NICU: Many NICU patients have heart activity monitored by electrocardiography (ECG), which can sample up to 1,000 readings a second to construct a waveform signal demonstrating the functioning of the heart. This translates to 86.4 million readings a day per patient. From this source signal, the ECG device also derives the heart rate and respiration rate, with each of these signals producing 86,400 readings a day per patient [4][8].

Drug and nutrition infusion data from smart infusion pumps (SIPs) contribute to the big data problem. SIPs can provide more than 60 different types of data every 10 seconds. If a new born stays in the NICU for 30 days, one SIP can generate 4.4 Mbytes of data per hour, 106 Mbytes of data per day, and 3 Gbytes of data monthly. Preterm infants can be connected to up to 13 SIPs, resulting in 39 Gbytes of drug infusion data from a single patient per month. And a lot of information could be generated and this information should be processed in order to help and survive new born children [6].

It's very important to mention that maternal health is closely linked to new born survival. While great strides have been made in reducing global child mortality, new-borns now account for 44 percent of all childhood deaths. Each year, 2.9 million new-borns needlessly die within their first month and an additional 2.6 million are stillborn. The main causes, which are preventable and treatable, are complications due to prematurity, complications during delivery, and infection [3].

The analysis showed that maternal smoking increased the relative risk of admission to an NICU by almost 20%. For infants admitted to the NICU, maternal smoking increased length of stay while for non-NICU infants it appeared to lower it. Over all births, however, smoking increased infant length of stay by 1.1%.

3.2 Healthcare Application

For this research, we will use 4 main characteristics to analyse: heart rate, respiratory rate, lower blood pressure and upper blood pressure. The data are generated randomly. We suppose that this data are generated by NICU. Diagram 1, diagram 2, diagram 3 and diagram 4 show how heart rate, respiration rate, lower blood pressure and upper blood pressure are being changed during the time.

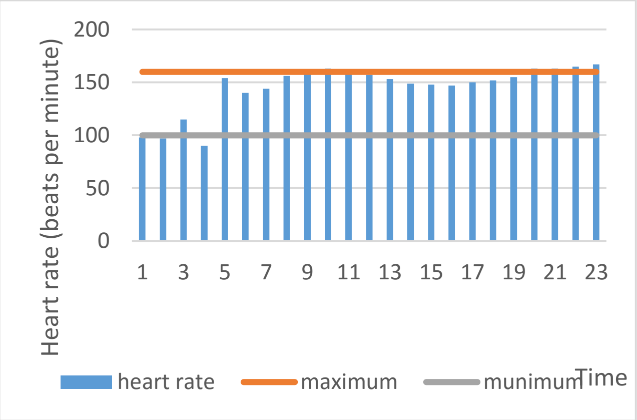

Diagram 1: Heart Rate

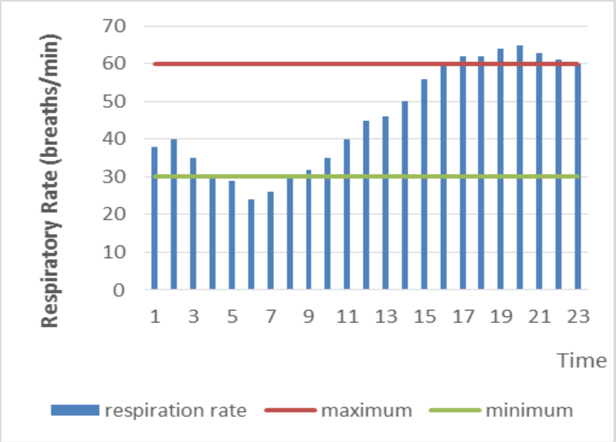

Diagram 2: Respiratory Rate

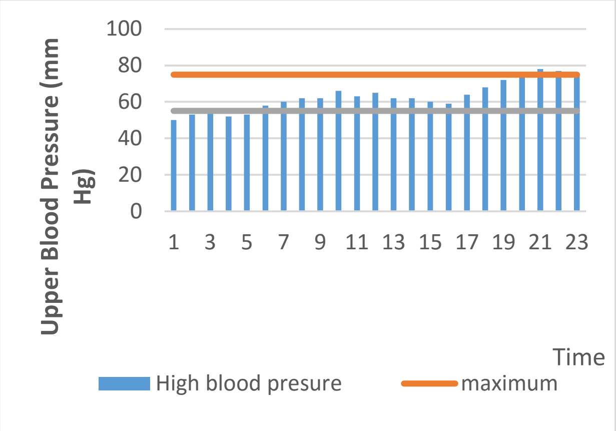

Diagram 3: Upper Blood Pressure (mm Hg)

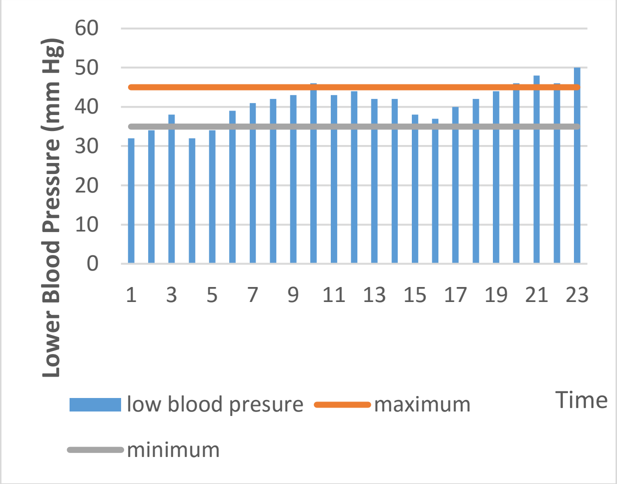

Diagram 4: Lower Blood Pressure (mm Hg)

Then we use Hadoop and Map Reduce programming in order to extract information for the patient or patients. A large part of the power of Map Reduce comes from its simplicity. Map Reduce works by breaking the processing into two phases: the map phase and the reduce phase [9][10][12]. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the map function and the reduce function. Below is presented the algorithm of simple program Word Count as well as one example of the demonstration of the program, which can be implied into the hospital.

3.3 Proposed Algorithm

The mapper emits an intermediate key-value pair for each word in a document.

The reducer sums-up all the counts for each word.

```
class Mapper
    method Map(docid a; doc d)
        for all term t ∈ doc d do
            Emit(term t; count )
```

```
class Reducer
    method Reduce(term t; counts [c1; c2; : : :])
        sum = 0
        for all count c ∈ counts [c1; c2; : : :] do
            sum  = sum + c
        Emit(term t; count sum)
```

To visualize the way the map works, consider the following sample lines of input data.

name0-96-43-27-22
name2-77-170-4-28
name1-75-80-103-21
name0-76-42-28-9
name2-74-188-77-30
name1-75-70-91-67
name0-69-173-80-36
name2-75-47-7-36
name2-87-101-23-32
name1-89-93-40-45
Etc...

Here the lines present records of the patients. Every part of the line has its own meaning: Thus, first part is the name; second part is the heart rate, then upper blood pressure and lower blood pressure and finally respiration rate.

These lines are presented to the map function as the key-value pairs:

(0, name0-96-43-27-22)
(16, name2-77-170-4-28)
(32, name1-75-80-103-21) Etc...

The keys are the line offsets within the file, which we ignore in our map function. The map function merely extracts the name and the heart rate (indicated in bold), and emits them as its output (the heart rate values have been interpreted as integers):

(name0, 96); (name2, 77); (name1, 75) etc...

The output from the map function is processed by the Map Reduce framework before being sent to the reduce function. This processing sorts and groups the key-value pairs by key. So, continuing the example, the reduce function sees the following input:

(name0, [96, 76, 69]); (name1, [75, 75, 89]); (name2, [77, 74, 75, 87])

Each name appears with a list of all its heart rate readings. All the reduce function has to do now is iterate through the list and pick up the maximum reading:

(name0, 96); (name1, 89); (name2, 87)

This is the final output: the maximum heart rate for each patient. The information can be saved as a txt file and will be ready for future use and analysis.

4. EXPERIMENTAL RESULTS

In this research, we used Intel® Core™ i5-660, 3.33GHz×4, Memory 3.7GB computer. In the file, which is going to be analysed, we have millions of lines. The size of the file is 37.9MB (1 million lines). This file contains 4 characteristics of 10 patients. After the running of the program, we can see that the map reduce program works correctly and emits data we need. In this case, we see the maximum heart rate for each patient. The same algorithm could be used to emit data about other characteristics (respiration rate, lower blood pressure etc.).

The doctor also can estimate the time. In this case, we see that the CPU time spent is 9480 ms. In order to decrease the time we only should add computers.

In addition to this we offer the equation below to evaluate the state of the patient:

$$F = \alpha X + \beta Y + \gamma Z + \theta Q$$

Where:

$$X = \frac{X_1}{X_{\max}}; \quad Y = \frac{Y_1}{Y_{\max}}; \quad Z = \frac{Z_1}{Z_{\max}}; \quad Q = \frac{Q_1}{Q_{\max}}$$

X_{max} – maximum heart rate of the patients

Y_{max} – maximum respiratory rate of the patients

Z_{max} – maximum upper blood pressure or the patients

Q_{max} – maximum lower blood pressure of the patients X_1, Y_1, Z_1 and Q_1 are current heart rate, respiratory rate, upper blood pressure and lower blood pressure representatively.

In the formula the sum of the coefficients (α, β, γ, θ) is 1. It will make the F formula give us a value between 0 and 1. In this paper we consider that α=β=γ=θ but later the exact values might be defined more precisely by working with IT, natural science and medicine researchers. As we have 4 characteristics (n=4) the coefficients can be evaluated in this way:

$$\alpha=\beta=\gamma=\theta=1/n;$$

So we will get that α=β=γ=θ=0.25.

In order to evaluate patients' condition we consider that:

$$F \in [0, 0.3] \rightarrow Bad\ condition$$

$$F \in [0.31, 0.8] \rightarrow Normal\ condition$$

$$F \in [0.81, 1] \rightarrow Dangerous\ condition$$

For example let us assume that X_1= 98, Y_1=38, Z_1=50 and Q_1=32; according to the database X_{max}= 200, Y_{max}=80, Z_{max}=100, Q_{max}=50.

$$F = 0.25 * \frac{98}{200} + 0.25 * \frac{38}{80} + 0.25 * \frac{50}{100} + 0.25 * \frac{32}{50} \approx 0.52$$

As F≈0.52 it means that the condition of the patient is normal.

5. CONCLUSION

Recently we have seen huge advances in the amount of data we generate and collect, as well as our ability to use technology to analyse and understand it. One of the most interesting field Big Data can occur is health care. Big Data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. To reach the goals doctors in the hospital and developers should work together to facilitate the creation of the platform as well as for privacy reasons when the information of the patients is collected. In this paper we have presented an application to deal with Big Data in health care. We have defined 4 variables related to patients. We have proposed an algorithm to analyse the values of these variables. Finally, it is presented an analytical formula to estimate the health of the patients depending on the current values of the variables.

6. REFERENCES

[1] McGregor, C, Catley, C., James, A., & Padbury, J. (2011). Next Generation Neonatal Health Informatics with Artemis. Medical Informatics Europe (MIE) 2011.

[2] Adams EK1, Miller VP, Ernst C, Nishimura BK, Melvin C, Merritt R. 2002, Neonatal health care costs related to smoking during pregnancy

[3] http://www.gatesfoundation.org/What-We-Do/Global-Development/Maternal-Newborn-and-Child-Health

[4] Carolyn McGregor, University of Ontario Institute of Technology, Canada. Big Data in Neonatal Intensive Care

[5] Bernard Marr, April 2015, Forbs. How Big Data Is Changing Healthcare. http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#6dcf69cf32d9

[6] C. McGregor et al., "Late Onset Neonatal Sepsis Detection in Newborn Infants via Multiple Physiological Streams," J. Critical Care, vol. 28, no. 1, 2013, pp. e11-e12.

[7] M. Blount et al., "Real-Time Analysis for Intensive Care: Development and Deployment of the Artemis Analytic System," IEEE Eng. in Medicine and Biology Magazine, vol. 2010, pg 110 – 118

[8] Carolyn McGregora, Andrew Jamesb,c,a, Mike Eklundd, Daby Sowe, Maria Eblinge, Marion Blounte © 2013 IMIA and IOS Press. Real-time Multidimensional Temporal Analysis of Complex High Volume Physiological Data Streams in the Neonatal Intensive Care Unit

[9] Hadoop: The Definitive Guide, Tom White, 2012

[10] Hadoop: Open source implementation of MapReduce. http://lucene.apache.org/

[11] http://bigdatauniversity.com/

[12] http://hadoop.apache.org