Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2017)

Towards of a Real-time Big Data Architecture to Intensive Care

André Gonçalves[a], Filipe Portela*[a], Manuel Filipe Santos[a] and Fernando Rua[b]

[a]*Algoritmi Research Center, University of Minho, Guimarães, Portugal*
[b]*Intensive Care Unit, Centro Hospitalar do Porto, Portugal*

Abstract

These days the exponential increase in the volume and variety of data stored by companies and organizations of various sectors of activity, has required to organizations the search for new solutions to improve their services and/or products, taking advantage of technological evolution. As a response to the inability of organizations to process large quantities and varieties of data, in the technological market, arise the Big Data. This emerging concept defined mainly by the volume, velocity and variety has evolved greatly in part by its ability to generate value for organizations in decision making. Currently, the health care sector is one of the five sectors of activity where the potential of Big Data growth most stands out. However, the way to go is still long and in fact there are few organizations, related to health care, that are taking advantage of the true potential of Big Data. The main target of this research is to produce a real-time Big Data architecture to the INTCare system, of the Centro Hospitalar do Porto, using the main open source big data solution, the Apache Hadoop. As a result of the first phase of this research we obtained a generic architecture who can be adopted by other Intensive Care Units.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Intensive Medicine; Intensive Care Units; Real-time; Big Data; Architecture; Hadoop.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: cfp@dsi.uminho.pt

Peer-review under responsibility of the Conference Program Chairs.

1. Introduction

Over the past 20 years, the amount of data has increased in large scale in several areas [1]. According to an International Data Corporation (IDC) report [2], in 2010, the amount of data created and replicated exceeded the Zettabytes barrier, reaching in 2011 the 1.8 Zettabytes. As at the date of the report, in 2011, the prospect was that this growth would increase nine-fold over the next five years [2], meaning that in 2016 was expected that the volume of data was close to 17 Zettabytes. According to an EMC report with research and analysis by IDC, health care accounts for a significant percentage of the data in the digital universe. In 2013, digital data about health care was 153 Exabytes, which, with an accelerated growth rate, in the order of 48% per year, is expected that the volume reach 2314 Exabytes in 2020 [3]. Yet, in the same report, they indicate that the global digital universe presents a growth of 40% a year, allowing to conclude that the growth of digital data, in health care, will be faster than the rest of the digital universe.

The constant, predictable and significant data increase has led to the introduction of a new paradigm, the Big Data. The term Big Data has come up with the explosive increase in data globally and is mainly used to describe huge data sets. In comparison to traditional data sets, Big Data is often associated with real-time analysis of large amounts of unstructured data [1]. At the level of health care, the information technology, the business world and the clinical research are creating an emerging movement under the Big Data banner [4]. Proponents of this movement argue that it is a new approach that will transform and accelerate health care, correcting decades of misguided research, and reshape clinical science as well as how to care for patients. Intensive Care Units (ICUs) are an environment in which the demand for change is increasing due to the constant recording and processing of large amounts of data, related with patients' health conditions [5]. According to Portela et al. [5], the complex condition of critically ill patients and the enormous amount of data, may hinder the decision-making, by intensivists, at the time of providing the best health conditions. The authors also add, that intensivists do not have the time to analyse the conditions of the patient in an assertive and consolidated way, and it is becoming increasingly necessary to develop systems capable of assisting intensivists in the shortest period, i.e., in real-time. In the first phase of this research, based on the INTCare architecture, other analysed big data architectures, and a superficial analysis of the Apache Hadoop ecosystem, we produced a Big Data architecture, which we believe can be used in any Intensive Care Unit. As a way of validating and release a final version of the architecture, to the INTCare, in the second phase of this research, we'll perform a detailed analysis of the Hadoop ecosystem and produce a prototype that will simulate the processing and storage of streaming data from bedside monitors, collected in the ICU of the Centro Hospitalar do Porto (CHP).

Finally, in addition to the introduction, this article is composed of four other sections. Second section, Background, provides background knowledge on topics related to Intensive Medicine and Intensive Care Units, the INTCare system, the main theme, Big Data and related work. Next, the third section identifies the research methodology used and how it applies in this research project. The fourth section describes the proposed real-time Big Data architecture, based on open source technologies. In the last section, the conclusions are presented, as well as the future work.

2. Background

2.1. Intensive Medicine, Intensive Care Units and INTCare Project

Intensive Medicine (IM) is a multidisciplinary area of medical sciences focused on the prevention, diagnosis and treatment of serious diseases, that are considered as threats to the lives of patients and which are characterized by causing failure of one or more vital organs [6]. Intensive Care Units (ICUs) are special hospital units prepared to provide health care to patients whose survival depends on intensive care. In these units, patients' vital signs are continuously monitored by various life support devices, which together with drug delivery enable patient recovery [6].

INTCare is a project developed at the Intensive Care Unit (ICU) of the Centro Hospitalar do Porto (CHP), which emerged from the need to build an intelligent system to automate the data collection and analysis process and, consequently, predict organ failure and their effects on patients [7]. The original system has suffered many changes, resulting from the expansion of ICU needs and the potentialities generated by the growing amount of electronic data. As a result, it is currently a Pervasive Intelligent Decision Support System (PIDSS) that acts automatically and in real-time, to provide more information to the intensivists of the UCI [7]. The INTCare system presents an architecture [8]

composed of four subsystems (data acquisition, knowledge management, inference and interface), that interact between themselves through the intelligent agents [8].

2.2. Big Data

Talking about Big Data is to approach an abstract concept[1]. Big Data is, in many respects, a poor term[9]. The concept has been used by the sciences to refer to data sets, large enough, to require supercomputers, but what was previously run by these machines can now be processed with standard computers and standard software[10]. However, even though Big Data is only associated with recording large amounts of data, its main features are the ability to search, aggregate and cross large datasets[9]. According to Chen et al.[1], at the present, although the importance of Big Data is generally recognized, there are still differences of opinion about its definition. The authors compiled various sources and, overall, the Big Data definitions converge on "a set of data that cannot be understood, acquired, managed and processed by traditional information technologies and software / hardware, within an acceptable period of time". This definition was first announced in 2010 by the Apache Hadoop team, responsible for one of Big Data's main open source projects, as the definition they considered valid. In fact, the Big Data concept was first approached in 2001, when a META analyst named Doug Laney presented the 3Vs model - Volume, Velocity and Variety - in a research report. This model was used by some research departments of Microsoft and IBM for the next 10 years [1].

2.3. Related Work

As a result of the realized research, we found the architecture of the Artemis platform (Fig. 1). It is an online platform for Neonatal Intensive Care Units, that enables simultaneous diagnosis of multiple patients, through real-time analysis of multiple data streams [11, 12].

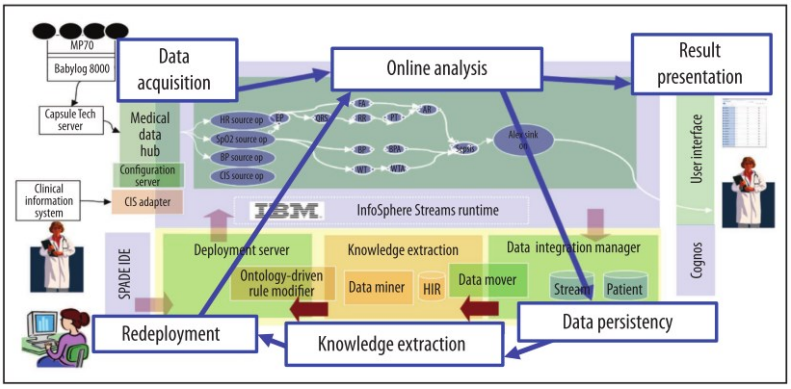

Fig. 1. Artemis platform architecture [12].

The Data acquisition component continuously inserts, into the platform, data streams from clinical devices and available clinical information. After that, the data is transferred to the Online analysis component, where the data is processed in real-time. In the Data persistence component, the data obtained in the Data acquisition component is stored together with the data generated in the Online analysis component. Subsequently, the data are exploited in the Knowledge extraction component and are specifically adapted to support clinical research for a set of conditions. Finally, the Redeployment component sends new, clinically-validated algorithms to the online analytics component[13].

3. Research Methodology

Despite the growing adoption of Big Data, there is still no method designed specifically for the definition and implementation of Big Data architectures / projects. After analysing the various methods of scientific research, we choose the Design Science Research Methodology for Information Systems. This method aims to study, research and

investigate the artefact and its behaviour from an academic or organizational point of view and it's composed by six activities, which constitutes the iteration process developed by Peffers et al. [14]. The first activity Identify problem & motivate seeks to define the problem to investigate and justify the value of the solution. Define objectives of a solution is the second activity and pursues to infer the objectives of a solution from the definition of the problem and from the knowledge of what is possible and feasible. Following in third, the activity Design & development is dedicated to creating the artefact resulting from the investigation, and may contain concepts, models, methods or instantiations. Demonstration is the fourth activity and consists in demonstrating the use of the artefact in the resolution of one or more instances of the problem, through experimentation, simulation and proof of concept. In the fifth activity, Evaluation, aims to observe and quantify / measure how well the artefact supports the solution to the problem by comparing the solution objectives with the results of the Demonstration. Lastly, the sixth activity, Communication, involves presenting the problem and its importance, the artefact, its usefulness, the rigor used in the project, and its effectiveness, to the researchers and professionals in the area. Within this research project, the iteration process will have a problem-centered initiation, once the problem and its importance are identified. In this sense, it was defined how each of these six activities relates to this specific research.

The problem here is the inability to process and store, in real-time, large amounts of data, coming from several different sources and platforms. The objective defined for this research seeks to find a solution and develop an artefact capable of processing and storing the data in real-time. The solution was to develop a real-time Big Data architecture based on open source technologies. The artefact in production is a Hadoop cluster, which will later be implemented for demonstration at the Intensive Care Unit (ICU) of the Centro Hospitalar do Porto (CHP). Once implemented, your performance will be evaluated by information systems administrators and by the intensivists who work there. Finally, the dissemination of this artefact and its importance will be done through a dissertation report, under work, and scientific articles, such as this one.

4. Real-time Big Data Architecture

The need to implement a Big Data architecture arose from the INTCare system's inability to process large amounts of data in real-time. Currently, the INTCare databases have an approximate size of 120 GB and the data is processed every 1 minute. For best results, INTCare needs to process the streaming data every 1 second. This requirement implies over annual processing of streaming data of 33 GB (when processed per minute) to 2 TB (when processed per second) and a total annual of 7 TB of stored data. The Real-time Big Data Architecture in Fig. 2 is a proposal of a generic architecture for intensive care, which is an evolution of INTCare architecture, capable of processing, storing, and analyzing large data sets, in real-time, using open source Big Data technologies, like Apache Hadoop. Thus, the architecture is composed of the stakeholders and five subsystems, composed by a set of intelligent agents.

4.1. Stakeholders

The stakeholders involved in the intensive care are: i) patients; ii) healthcare professionals (doctors, nurses …); iii) government and administration (those who choose and buy services and technology from the service and solution providers); iv) healthcare service providers; v) data scientists (people who analyse the data, create scenarios …); and vi) technology solution providers (people who provide the healthcare technology). These *stakeholders* may be users or data providers.

4.2. Data Acquisition, Data Management and Knowledge Management

Data Acquisition subsystem is responsible for the acquisition of data from several different sources, such as: external databases, documents, images, sensor and others. Depending on the specificity of each of the data sources, there may be one or more intelligent agents in the data acquisition, which will send the data to the interface and data management subsystems. The Data Management represents a Hadoop cluster, where the large data sets are processed and stored. The Hadoop components used in this subsystem depends on the type of data and the requirements of the environment. Knowledge subsystem can contain intelligent agents for: i) Data Mining (that converts data recorded in the data warehouse into knowledge through the creation of real-time forecast models); ii) Data Analytics (integrated

with other applications to provide examination of the large data sets); iii) Performance (for the statistical data collection); and iv) Ensemble (to combine several models with the goal of improving predictive performance).

4.3. Inference and Interface

Integrated with Interface subsystem, and used by the healthcare professionals and data scientists, the Inference subsystem can provide two intelligent agents: Prediction and Scenario Evaluation. The Prediction Agent uses the data, produced by the application of the models implemented in the Knowledge Management subsystem, to answer questions from users. The Scenario Evaluation Agent provides to the users the capability to create and evaluate hypothetical scenarios. Composed of an intelligent agent, this subsystem provides a web interface that integrates with other subsystems, allowing the healthcare professionals to access the most diverse information of patients, evaluate scenarios, request prognoses, and others.

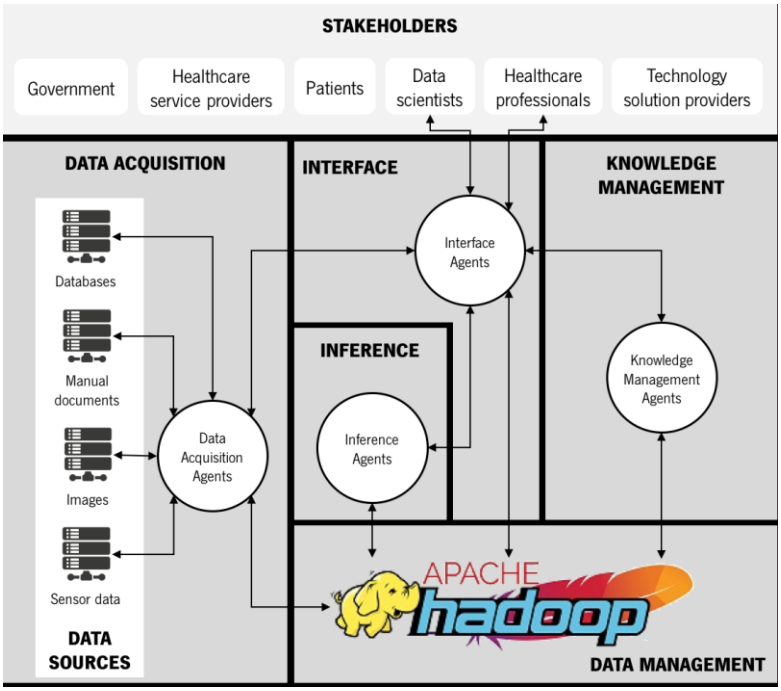

Fig. 2. Real-time Big Data Architecture for Intensive Care.

4.4. Intelligent Agents

This type of agents makes the system work through automatic actions, which perform some essential tasks, such as automatic data collection and updating of predictive models, in real-time, without the need of human intervention [15]. The interaction between the agents is a focal point in the efficiency and flexibility of the system [16]. The Real-time Big Data architecture has for agents: Data Acquisition, Inference, Interface and Knowledge Management.

5. Conclusions and Future Work

This research provided useful information about how the data has been growing at a very fast away and its impact and potential in the health care sector. The inability to process and store large amounts of data, in real-time, requires the implementation of Big Data solutions in order to gain competitive advantage and extract, increasingly, value from data. The present INTCare system is a clear example of a system whose need to implement a Big Data architecture is

critical. The main limitation is the inability to process streaming data with a recurrence less than 1 minute, given that it would go from 33 GB of processed data, per year, to 2 TB (with a recurrence of 1 second).

From the analysis of the architecture of the Artemis platform, we understand that adding a fifth subsystem to the present architecture of the INTCare, we obtain a solution capable of responding to the needs of processing and storage, of the large amounts of data. In addition, this new architecture can be used in any Intensive Care Unit that wants to implement a solution based only in Big Data open source products.

The architecture designed is the main contribution of this work. This architecture represents a global definition of the Intensive Care requirements and it can be deployed in any services which has similarity of specifications, i.e., real-time data acquisition and processing needs - critical services with reduce time to make decisions.

As future work, it is necessary to carry out an analysis of the needs of the INTCare system and a detailed identification of all projects of the ecosystem of the Hadoop project, in order of selecting the right Hadoop projects and design a specific architecture for the INTCare. A proof of concept is also defined to assessing the architecture feature. After that, and before the implementing and testing the architecture, it should be performed a predictive analysis of the hardware requirements, with the purpose of the scaling the Hadoop cluster.

Acknowledges

"This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013." This work is also supported by the Deus ex Machina (DEM): Symbiotic technology for societal efficiency gains - NORTE-01-0145-FEDER-000026

References

1. Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications, 19*(2), 171–209.
2. Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos. *IDC iView*, pp. 1–12.
3. EMC with Research & Analysis by IDC. (2014). The Digital Universe Driving Data Growth in Healthcare. Retrieved December 29, 2015, from https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf.
4. Iwashyna, T. J., & Liu, V. (2014). What's so different about big data? A primer for clinicians trained to think epidemiologically. *Annals of the American Thoracic Society, 11*(7), 1130–5.
5. Portela, F., Santos, M., Vilas-Boas, M., Rua, F., Silva, Á., & Neves, J. (2010). *Real-time Intelligent decision support in intensive medicine*. Paper presented at the KMIS 2010- International Conference on Knowledge Management and Information Sharing.
6. Filipe Portela, Manuel Filipe Santos, José Machado, António Abelha, Álvaro Silva and Fernando Rua Martins. Step towards Pervasive Technology Assessment in Intensive Medicine. International Journal of Reliable and Quality E-Healthcare (IJRQEH). Volume 6, Issue 2, pp 1-16. ISSN: 2160-9551. IGI Global. (2017). DOI: 10.4018/IJRQEH.2017040101
7. Portela, F., Santos, M., Machado, J., Abelha, A., Silva, Á., & Rua, F. (2014). Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture. In M. Bursa, S. Khuri, & M. E. Renda (Eds.), *Information Technology in Bio- and Medical Informatics SE - 9* (Vol. 8649, pp. 87– 102). Springer International Publishing.
8. Santos, M.F., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., Neves, J., Silva, A., Rua, F.. A Pervasive Approach to a Real-Time Intelligent Decision Support System in Intensive Medicine. CCIS - Communications in Computer and Information Science. Volume 272, 2013, pp 368-381. ISBN: 978-3-642-29763-2. Springer. (2012). doi.org/10.1007/978-3-642-29764-9_25
9. Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA. *Information, Communication & Society, 15*(5), 662–679.
10. Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. Retrieved from http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big- social-data/64-article-2011.pdf
11. Blount, M., Ebling, M. R., Eklund, J. M., James, A. G., McGregor, C., Percival, N., ... Sow, D. (2010). Real-time analysis for intensive care: development and deployment of the artemis analytic system. *IEEE Engineering in Medicine and Biology Magazine: The Quarterly Magazine of the Engineering in Medicine & Biology Society, 29*(2), 110–8.
12. McGregor, C., Catley, C., James, A., & Padbury, J. (2011). Next generation neonatal health informatics with Artemis. *Studies in Health Technology and Informatics, 169*, 115–9.
13. McGregor, C. (2013). Big Data in Neonatal Intensive Care. *Computer, 46*(6), 54–59.
14. Peffers, K., Tuunanen, T., Rothenberger, M. a., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24*(3), 45–78.
15. Santos, M. F., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., & Neves, J. (2011). *INTCARE - Multi-agent approach for real-time Intelligent Decision Support in Intensive Medicine*. Paper presented at the 3rd International Conference on Agents and Artificial Intelligence (*ICAART*).
16. Gago, P., Santos, M. F., Silva, Á., Cortez, P., Neves, J., & Gomes, L. (2006). INTCare: a knowledge discovery based intelligent decision support system for intensive care medicine. *Journal of Decision Systems*.