# How to Manage Efficiently Clinical Big-Data by Means of Cloud Computing

Antonino Galletta(✉) and Massimo Villari

MIFT Department, University of Messina, Messina, Italy
{angalletta,mvillari}@unime.it

**Abstract.** Nowadays, Information and Communication Technologies (ICT) are widely adopted in hospitals. Increasingly often medical devices are computer-assisted. Hospital Information Systems (HISs) are not designed to manage the huge amount of data produced by these devices. New paradigms, such as Cloud Computing, by means of its features represents a valid tool to handle this kind of problem. Cloud Computing is very powerful, but it arises issues concerning data privacy. For this reason, clinical operators are reluctant to adopt it in HISs. In this paper, considering two real use-cases coming from the IRCCS "Bonino Pulejo", a clinical and research center in Messina, we discuss a Cloud Computing architecture able to manage amounts vast of medical data. From a technical point of view, the proposed solution is based on microservices each of them realized for performing a specified task, such as the anonymizer. A microservice that is able to obfuscate users' sensitive data in order to assure data privacy and to make the system compliant with GDPR.

**Keywords:** Telemedicine · GDPR · Cloud computing · MRI · Internet of Things · Big data

## 1 Introduction

Nowadays, we are observing a revolution in hospital and clinical centers. Indeed, often old medical devices are replaced by innovative computer assisted ones. These new kind of equipment widely adopt Internet of things (IoT) approach. Statista [1] predicted that the number of connected devices will grow up to about 75 billion in 2025. These new kind of devices allow physicians to make more accurate and precise diagnoses. However, they produce a huge amount of data. These data are different for type and structure, therefore, the way to manage them is also different. Considering as example Magnetic Resonance (MR). It produces Digital Imaging and COmmunications in Medicine DICOM files: series of jpg images with a specific header. Instead equipments for rehabilitation such as CAREN and Lokomat produce raw data that can be stored in internal Hard-Disks or exported as Comma-separated values (CSV) files. Traditional

M. Villari—Supervisor.


M. Fazio and W. Zimmermann (Eds.): ESOCC 2018 Workshops, CCIS 1115, pp. 148–157, 2020.
https://doi.org/10.1007/978-3-030-63161-1_12

Hospital Information Systems are not able to manage these data. Innovations in ICT provide a very powerful tool suitable for solving such a problem: Cloud Computing. Use Cloud Computing in HIS provides several advantages for clinical centers. Indeed, it allows to create high available specific workflows that can scale up or down based on the workload. However, it presents several issues related to users' data privacy especially in the GDPR era. Such a problem makes difficult the proliferation of Cloud based systems in HIS. In this paper, we discuss about of the experience done during the first two years of the doctorate course at the University of Messina and the IRCCS Centro Neurolesi "Bonino Pulejo". In particular, considering two real medical use-cases coming from the IRCCS, one related to Magnetic Resonance (MR) and another related to rehabilitation, we discuss about of a Cloud Computing microservice architecture able to manage the huge amount of produced data. In our solution we widely adopt the Hybrid Cloud approach: sensitive data are stored in a secure manner into the Private Cloud, data that have to be shared with foreign users are stored into the Public Cloud. From a technical point of view, the proposed solution is based on microservices each of them realized for performing a specified task, such as the anonymizer. A microservice that is able to obfuscate users' sensitive data in order to assure data privacy and to make the system compliant with GDPR. The rest of the paper is organized as follows. In Sect. 2, we present other works related to telemedicine, BigData visualization and mechanisms to share data over the Cloud. In Sect. 3 we focus on our motivations for this work. The Sect. 4, describes the architecture designed for managing effectively clinical Big Data. Finally, in Sect. 6 we conclude the work discussing about of future steps.

## 2 Background and Related Work

Nowadays, ICT is present in all fields from the industry to the agriculture to the health care. Innovations in ICT word led to creation to new systems and protocols. Such as telemedicine, physicians by using it are able to provide assistance to remote bedridden patients. Benefits and drawbacks of telemedicine were discussed in [2].

The consequence of the introduction of these new systems was an explosion of data. The management of these data is very complicated, indeed traditional techniques and systems are not adequate to manage them. Another problem in Big Data is the analysis of them in order to extract insight.

In [3], authors presented a way for processing electroencephalography data. Their work is based on three steps: in the first step they convert data from European Data Format (EDF) to the JavaScript Object Notation (JSON); in the second they gather JSON data; in the latter, by means of smartphones, they perform real-time interactions with signal data.

A tool able to show the relationship among heterogeneous data is presented in [4]. This tool, based on three data structures (Tree Structure, Graph Structure and Graph-Tree Structure), shows the relationship of data stored into relational databases.

4 visualization tools for physicians were discussed [5]. These visualization tools show the behavior of measured parameters considering different time interval: the "Continuous Month" which groups measurement by month representing them day by day; the "Continuous Day" which groups measurement by day representing them hour by hour; the "Circular Day" which represents by means of a pie chart same parameters of the "Continuous Day"; the "Multi-Circular Day" tool, instead, allows to compare the behavior of specified parameter over several days.

In Big Data another challenge is the sharing of them among several users.

In [6], authors considering a Network Storage Environment (NSE) discussed a file partitioning method optimization. In particular, considering serviceability, reliability and availability, they proposed an algorithm for distributing files inside a cluster.

An approach for improving the file reliability is discussed in [7]. In particular, authors split data into chunks. The main idea is to increase the reliability of the system adding redundancy. In such a way, they can assure by means of data correction procedures. In PRESIDIO [8], is discussed a similar strategy.

Authors in [9] discussed a file partitioning strategy. In particular, they presented BerryStore: a distributed object storage system designed for Cloud service especially for the massive small files storing. By means of a distributed coordinated controller, BerryStore is able to assure concurrency, scalability, and fault-tolerance.

The management of Big Data is a very complex task. The common idea of all aforementioned scientific works is to create a specif tool per each type of data. In this work, we aim to create a single architecture that can be specialized based on the type of data.

## 3 Motivation

Nowadays, Information and Communication Technologies (ICT) are widely adopted in hospitals. Indeed more often medical devices are computer-assisted (e.g., Magnetic Resonance, Lokomat, CAREN). Data produced from these devices are different for dimension, form and quantity. For instance, considering the Magnetic Resonance (MR) it produces Magnetic Resonance Images (MRIs): series of jpg images with a specific header. Usually, these images are stored and processed as DICOM files. CAREN and Lokomat produce raw data related to rehabilitation activities of patients, these data are stored into internal Databases (DBs) and can be exported as CSV files. The difference among the structure of these kind of data reflects also differences in the management. Indeed in the first case the object storage is needed.

The research activities presented in this paper can be divided into two main branches: the first one related to the management of MRIs, the latter related to the management of rehabilitation data coming from CAREN, Lokomat and wearables. With reference to MRI branch, physicians need a system that allow them to share DICOMs in a safe manner with other practitioners, physicians and patients.

Share content among different practitioners is very important for a clinical point of view, indeed, it allows to merge the experience of different physicians and to have more accurate diagnosis. Instead, with reference to rehabilitative data, physicians need a telemedicine Big-Data visualization tool that allows them to analyse patients' health status simply looking at data representation.

The enabler technology that allows us to create a system able to manage effectively these kinds data is the Cloud Computing. In particular, in order to reach our goals, we adopted a Hybrid Cloud Computing approach. We used public Cloud Storage providers for storing anonymized DICOM images and Private Cloud in order to save personal and rehabilitative data. Public Cloud may arise many treats in terms of data security, privacy and availability. Indeed, Cloud Storage services might discontinued (such as copy https://copy.com/), or attacked by hackers (such as Dropbox. Indeed in 2012 68 Million of account was compromised [10]). If we also consider the GDPR the scenario become very complex. The end users of our system, physicians and patients, often are not accustomed to use this kind of tools, therefore they need an user friendly tool that allows them to manage these data. Considering the assumptions that we made before, we aim to realize an user friendly tool that allows physicians to manage clinical data in a secure manner.

The design of our Cloud based clinical Big Data management solution have to satisfy the following requirements:

1. single core that can be specialized based on users' requirements;
2. high scalability: capability to adapt the execution of different services to the workload;
3. compliance with GDPR;
4. ability to manage both real time acquisitions and historical data coming from different data sources;
5. user friendly interface suitable for Personal Computers and mobiles.

## 4 Our Approaches

In this Section, we discuss about of approaches adopted in order to manage effectively medical data. In particular, considering requirements described in the Sect. 3, we designed and developed two specific software prototypes based on microservices. The design of these prototypes starts from the same core that can be specialized based on users' requirements. We decided to adopt a microservice-based architecture in order to fulfill the scalability requirement. Indeed, each microservice can be migrated based on the workload from a machine with lower computation capabilities to a more powerful one.

In the following Subsection we discuss about of the design of these prototypes.

### 4.1 Big MRI Share

In this Subsection we discuss about of the software prototype designed for managing MRIs and share them with practitioners, physicians and patients.

Figure 1, shows whole architecture of the system for managing and sharing MRIs. The system is compound of 8 blocks:

1. Magnetic Resonance (MR) the source of MRIs;
2. OwnCloud, the Private Cloud adopted for storing MRIs;
3. Anonymizer, the GDPR compliant microservice that anonymizes sensitive patient's data;
4. MongoDB a Document Oriented Big Data database used as system database;
5. Splitter, the microservice that executes the data decomposition algorithm and spread data chunks over the Public Cloud Storage providers;
6. Public Cloud Storage providers, the public repository;
7. Meteor based app, a web-app that executes the data recomposition algorithm and displays MRIs to end users;
8. practitioners, the consumers of MRIs.

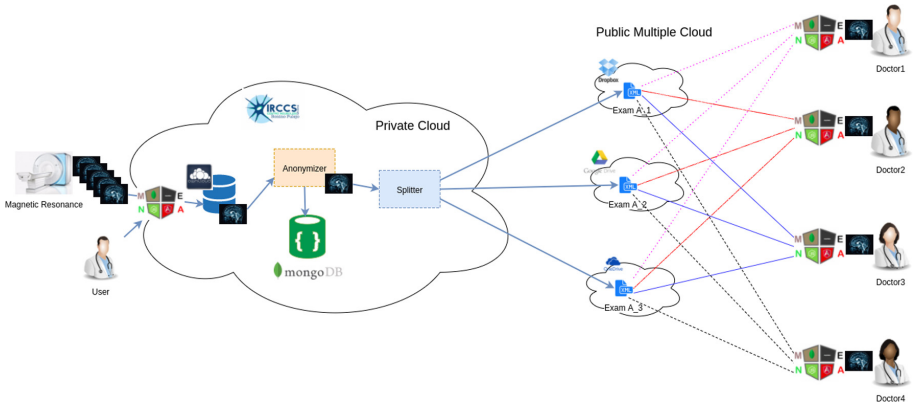


**Fig. 1.** Big MRI sharing architecture

Producer and Consumers generate and analyse DICOM files containing MRIs. With reference to the architecture showed above, the producer is the MR. In our system we could have two kind of consumers that need to analyse MRIs: foreign and internal practitioners. Each DICOM file is composed of thousands of images that are organized in series. In our system, we store each series into a specific OwnCloud directory. At this level, the data privacy is guaranteed by OwnCloud, indeed only authorized users can gain the access to the stored files.

External practitioners, do not have any way to access directly to data stored inside the Private Cloud. Only internal authorized physicians can share contents for a limited time period.

As discussed in Sect. 3, share clinical sensitive data on Public Cloud Storage services arises several privacy threats. We remark that one of the requirement is the compliance with GDPR. In order to satisfy this requirement and increase the

security of the whole system we created a specific microservice able to anonymize users' sensitive data.

The anomymization process updates the sensitive data contained into the header of DICOM files and store metadata inside MongoDB. More specifically this process updates the name of the patient, the user ID and the date of birth. The patient name is replaced by an UUID that depends on the DICOM series; the user ID is updated with a random number; regarding the date of birth the algorithm updates only day and month because the year could be useful in order to make diagnosis.

Anonymized DICOM files, from Anonymizer are sent to the Splitter microservice. It, by means of splitting algorithms such as the RRNS [11], divides the original file into chunks and spread them over the Public Cloud Storage services. The exact location of each chunk is stored into a Map-File, a special XML file composed of two main nodes: the header, that contains metadata (such as hospital and practitioner) and the data node that contains public paths to anonymized DICOM chunks.

The Map-File is very important during the recomposition phase. Indeed, it is passed as input to the Meteor web-app that runs the recomposition algorithm. The Meteor web-app represent the interface for external practitioners, it provides different functionalities such as DICOM visualization, identification and display of Region of Interest ROI etc.

### 4.2 Big Rehabilitative Data Visualization

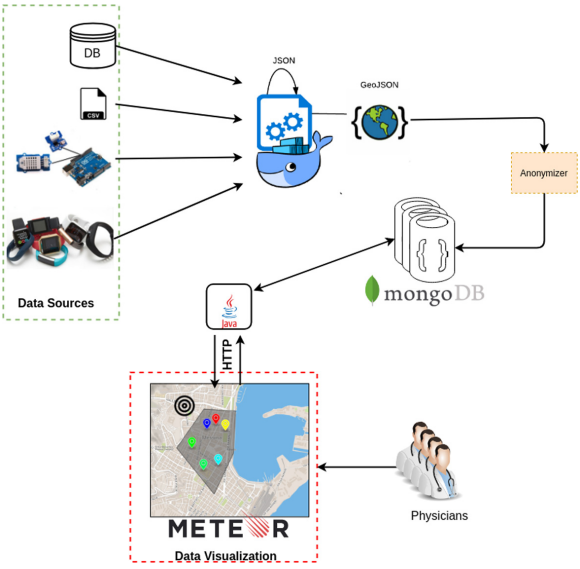


**Fig. 2.** Big rehabilitative data visualization architecture

In Fig. 2, is shown the overall architecture of the system able to manage rehabilitative data. Also in this case the architecture is based on microservices, in such a way several component, such as Anonimyzer and MongoDB, can be shared among different solutions.

The system is composed of seven blocks:

1. Data sources such as CAREN, Lokomat and wearable;
2. GeoJSON converter, the microservice that uniforms and transforms incoming data to GeoJSON;
3. Anonymizer, the GDPR compliant microservice that anonymizes sensitive patient's data;
4. MongoDB a Document Oriented Big Data database used as system database;
5. Mongo Interface, the microservice that acts as interface between front-end and back-end;
6. Meteor based app, a web-app that shows charts related to patients' data;
7. Practitioners, the consumers of data.

In our system we could have different data sources such as CAREN Lokomat and Wearables. Data produced from these sources can be stored into specific files or gathered in a real time fashion. We remark that the fourth requirement discussed into the Sect. 3 is the ability to manage both real time acquisitions and historical data coming from different data sources. In order to fulfill it, we created a microservice that by means of specific interfaces is able to interact with different data sources. Data acquired from these sources will be converted in GeoJSON, a standard for encoding geographic data structures.

We adopted this format because it is natively stored inside the database system, therefore we can make queries in a simple way.

The fifth requirement is related to the user friendliness of the system. In order to satisfy it, we created a Meteor web-app that shows data for physicians-defined geographical zones in two modalities: general overview or patient-specific [12].

For security concerns, MongoDB is not directly exposed to the external word. Thus we created a specific microservice that act as interface. It runs a Java application that is able to make query on MongoDB by means of the official MongoDB drivers and to receive command from the Meteor web-app.

## 5 Highlights and Discussions

In this Section we analysed our system from a numerical point of view. In order to validate the system, we made specific analyzes for each proposed approach. Our analyses can be divided into three categories: common aspect of presented architectures (scalability analysis of the anonymization process), MRI (in term of disk usage) and rehabilitative data (in term of time needed to make queries).

Our testbed is composed of microservices running on a web server with the following hw/sw characteristics: CPU Intel(R) Core(TM) i5-5200U CPU @ 2.8 8 GHz with 2 cores and 2 threads, RAM 8 GB, GFLOPS 66, OS: Ubuntu server

16.04 LTS 64 BIT. In order to have more reliable results we performed 30 consecutive iterations and considered confidence at 95%. In Fig. 3 the behavior of the anonymizer is shown. As the reader can observe the system scales up linearly with the increasing of the number of processed elements.

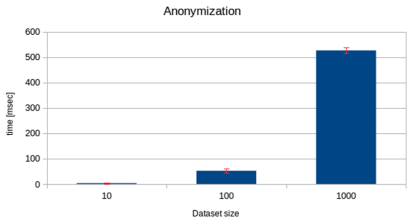


**Fig. 3.** Performance analysis of the anonymization module considering increasing dataset sizes.

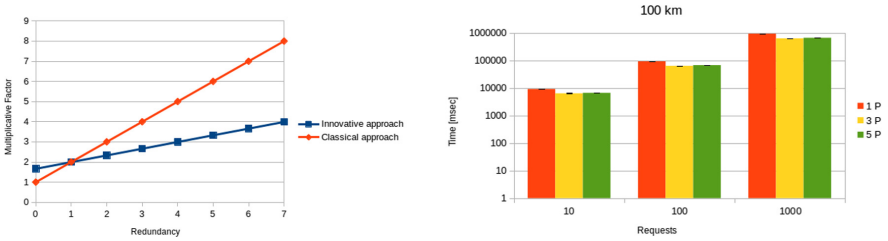


**Fig. 4.** Performance analysis of MRI management (a) and rehabilitative data (b)

In Fig. 4(a) is shown a comparison, in term of disk usage for store MRIs, between RAID 1 approach and RRNS. As the reader can observe, the capacity of the disk required from RAID 1 approach increase linearly with the redundancy. RRNS, instead, scales following a different behavior. Indeed, considering the case without redundancy it is less convenient of RAID 1, but considering the case with 7 degrees of redundancy it is more powerful, indeed it reduces the space required of a factor 1.75. For further the reader can refer to [13] and [14].

Considering the Big rehabilitative data, hereby we analyse performance of the general overview visualization mode. Our testbed is composed of 400k on random patients stored inside a specific MongoDB collection. As reference case, we considered a circular area of interest of 100 km. In our analyses, we considered three different configurations. In order to analyse the robustness of the system, we made 10, 100 and 1000 subsequent requests. Figure 4(b), show the behavior of the deployed system. As the reader can observe, time performances increase linearly with the number of subsequent requests that we performed.

Requests with a single patient's parameter are the simplest but time performance are slower. This behavior is due to the huge amount of data that

flows from MongoDB to practitioners. Requests with five patient's parameters have intermediate performance, indeed they present the more complex query but return back less results. From a numerical point of view, the better trade off is implemented by the request with three parameters, indeed it presents computation time lower than other scenarios. For further the reader can refer to [12].

## 6 Conclusions and Future Work

In this scientific work, we discussed about of the management of clinical Big-Data. In particular, considering two real use-cases one related to MRI and another one related to the rehabilitation, that were defined from the IRCCS Centro Neurolesi "Bonino Pulejo" of Messina, we described a Cloud based software architecture. During the design of this architecture, we considered five requirements such as i) the presence of a single core that can be specialized based on users' requirements; ii) high scalability of the system; iii) compliance with GDPR; iv) ability to manage both real time acquisitions and historical data coming from different data sources; v) user friendliness interface.

The architecture that fulfills above described requirements is based on microservices, each of them with a specific function such as the such as the anonymizer. A microservice that is able to obfuscate users' sensitive data in order to assure data privacy and to make the system compliant with GDPR.

In this report we discussed about of the experience done during the first two years of the doctorate course at the University of Messina and the IRCCS Centro Neurolesi "Bonino Pulejo". For the last year the plan is to spend six month at Karlstads University in order to work on the design of SDN-based geologically distributed solutions for Big Data analytics.

## References

1. statista: Iot number of connected devices worldwide. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/. Accessed Jan 2018
2. Hjelm, N.M.: Benefits and drawbacks of telemedicine. J. Telemed. Telecare **11**, 60–70 (2005)
3. Serhani, M.A., Menshawy, M.E., Benharref, A., Harous, S., Navaz, A.N.: New algorithms for processing time-series big EEG data within mobile health monitoring systems. Comput. Methods Programs Biomed. **149**, 79–94 (2017)
4. Liu, Q., Guo, X., Fan, H., Zhu, H.: A novel data visualization approach and scheme for supporting heterogeneous data. In: 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 1259–1263 (2017)
5. Frink, T.M., Gyllinsky, J.V., Mankodiya, K.: Visualization of multidimensional clinical data from wearables on the web and on apps. In: 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), pp. 1–4 (2017)

6. Hai-Jia, W., Peng, L., Wei-wei, C.: The optimization theory of file partition in network storage environment. In: 2010 9th International Conference on Grid and Cooperative Computing (GCC), pp. 30–33 (2010)
7. Bhagwat, D., Pollack, K., Long, D.D.E., Schwarz, T., Miller, E.L., Paris, J.F.: Providing high reliability in a minimum redundancy archival storage system. In: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation. MASCOTS 2006, Washington, DC, USA, pp. 413–421. IEEE Computer Society (2006)
8. You, L.L., Pollack, K.T., Long, D.D.E., Gopinath, K.: Presidio: a framework for efficient archival data storage. Trans. Storage **7**, 6::1–6:60 (2011)
9. Fan, K., Zhao, L., Shen, X., Li, H., Yang, Y.: Smart-blocking file storage method in cloud computing. In: 2012 1st IEEE International Conference on Communications in China (ICCC), pp. 57–62 (2012)
10. BBC: Dropbox hack' affected 68 million users'. https://www.bbc.com/news/technology-37232635. Accessed 30 July 2018
11. Celesti, A., Fazio, M., Villari, M., Puliafito, A.: Adding long-term availability, obfuscation, and encryption to multi-cloud storage systems. J. Netw. Comput. Appl. **59**, 208–218 (2016)
12. Galletta, A., Carnevale, L., Bramanti, A., Fazio, M.: An innovative methodology for big data visualization for telemedicine. IEEE Trans. Ind. Inf. **15**(1), 490–497 (2018)
13. Galletta, A., Celesti, A., Tusa, F., Fazio, M., Bramanti, P., Villari, M.: Big MRI data dissemination and retrieval in a multi-cloud hospital storage system. In: DH (2017)
14. Galletta, A., Bonanno, L., Celesti, A., Marino, S., Bramanti, P., Villari, M.: An approach to share MRI data over the cloud preserving patients' privacy. In: 2017 IEEE Symposium on Computers and Communications (ISCC), pp. 94–99 (2017)