A Research Infrastructure for E-Health Big Data Analytics

Alessio Botta[1], Elio Masciari[1]

[1]University of Napoli Federico II, Napoli, Italy
alessio.botta@unina.it
[1]University of Napoli Federico II, Napoli, Italy
elio.masciari@unina.it

Abstract

In this paper, we present a research infrastructure, built within the Department of Excellence project in order to support a wide spectrum of big data analysis related to e-health. More in details we built a both a public cloud based infrastructure and a private cloud one in order to guarantee a high performance approach to researchers.

Keywords

Big Data Platform, Cloud Services for Big Data, Research Infrastructures.

1. Introduction

The Department of Electrical Engineering and Information Technologies - DIETI - of the University of Naples Federico II is the largest department in Southern Italy that works on issues relating to Information and Communication Technology (ICT). The scientific/disciplinary sectors (Settori Scientifico/Disciplinari in Italian or simply SSD) participating in the activities of DIETI have shown in the last round of Evaluation of the quality of research (VQR) 2011-2014 evaluation peaks of absolute excellence; very limited is in fact the number of SSD below the national average evaluation. With particular reference to the Area 09, ten SSD have received evaluations greater than or equal to the national average. The DIETI, unanimously, has therefore decided to develop in the next five years its research lines in the field of Information and Communication Technologies, especially with regard to the application of modern information technologies in the thematic areas of the so-called eHealth. From these premises it was natural to propose an innovative project called *ICT for Health* (ICTH in the following) within the Departments of Excellence call of the Italian Ministry of Research and University. The call financed 180 Excellence Departments in Italy, and DIETI was among the 14 ones with the maximum evaluation for the project, one of the only two cases in south Italy. ICTH has been financed with over 9 M Euros for a time span of five years.

Using the financial tools available for the project, DIETI is recruiting a full professor chosen by a panel of international experts from a short-list of candidates of excellence. This figure is responsible for coordinating and managing the entire project ensuring the success and

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

achievement of all objectives. In addition, it is currently ongoing the recruitment of an associate professor and four Tenure-track Researchers (RTD-B) in order to strengthen the research areas that are described below. The ICT for Health project involves the strengthening of existing laboratories in the DIETI and the creation of a new experimental reality active on issues not present now within the DIETI, in order to create a network of laboratories to support the proposed research. Finally, a new doctorate school has been created, specific on the themes of the project in order to have, together with the development of the research activities, also the training of professionals ready to be employed in a socio-economic regional and national environment, particularly active in this field and whose annual European budget now exceeds 20 billion euros. In the following we will describe the requirements we would like to fulfill with our Big Data laboratory.

1.1. Research Agenda

The eHealth technologies and services in the coming decades will bring deep changes in the organization of the health care system, with the aim of improving the quality and efficiency of care and, at the same time, reducing costs. In the short term, the new technologies will have to integrate with the current structures that provide healthcare, but in the long term they will produce significant changes both in the internal organization and in the architecture of the buildings that will house the hospitals and healthcare facilities of the future. These facilities will need to be able to provide personalized therapies to patients, which may also be decentralized or delivered at home and monitored remotely. The ICT for Health departmental project, based on the pre-existing infrastructure of the DIETI laboratories and on the new laboratory to be built, is therefore developed along the following lines of scientific and technological research.

1.1.1. Sensing for Health

Smart transducers for the Internet of Everything environment will be developed, such as the electronic syringe, smart sensors to integrate robotic microsurgery devices, advanced biomedical instrumentation for automated and robotic surgery systems, and innovative systems for remote monitoring of patients' health status, both in hospitals or healthcare facilities (e.g., nursing homes for the elderly), and in private homes.

In particular, we will study solutions based on new sensors and wearable devices, wearable or implantable, which will allow to monitor the state of health even outside the traditional care centers. Particular attention will be paid to Exergaming technologies based on Serious Games, in which rehabilitation or treatment takes place through Augmented Reality applications, and Brain Computer Interfaces, with low training and response times, and low number of electrodes, for patients with high rates of disability. The topics of sensor development and their interaction in augmented vision systems will be experimentally developed within the new laboratory foreseen in the program.

1.1.2. Data for Health

The eHealth services and technologies generate huge amounts of data and information that, for their treatment, require the use of non-traditional methodologies and technologies: Cloud

Computing, Internet of Things and Big Data Analytics are the new paradigms underlying the new generation of systems for information management in eHealth. The data sources to be considered, in addition to having high volume, are also heterogeneous given their different types and origins [1]. The speed with which information is produced and stored, together with the aforementioned volume and variety, require systems and tools to collect, manage, and analyze the data and information produced by healthcare systems, directing research towards Big Data Analytics (BDA) methodologies and techniques [2].

The BDA in eHealth enables the transformation of a classical hypothesis-driven information analysis to an innovative data-driven one, able to identify non-trivial connections between heterogeneous data and information. This requires the need to investigate: (a) new cloud-based architectures that allow for the timely processing of information, from the Hadoop to the Spark Ecosystem; (b) new information management systems that integrate relational (SQL), non-relational (NoSQL) and new relational (newSQL) architectures; (c) use of descriptive, diagnostic, prescriptive and descriptive analysis techniques; (d) use of Data Mining tools and techniques on Massive Data Sets, including Deep Learning . In this context, equally important are digital infrastructures for data circulation and device interconnection, following the Internet of Things paradigm.

1.1.3. Logistics for Health

In-hospital logistics services impact 15-20% of operating costs. These services include moving patients, linen, meals, medications, equipment and samples between clinics, wards, operating rooms, laboratories and warehouses. Digitization, automation, and robotic technologies can optimize these processes through solutions that enable automated patient and material transport and automated hospital warehouse management. The main difference from factory logistics systems is the need to operate in man-made environments.

We aim at creating robotic systems capable of interacting with humans (patients, medical/nursing staff, visiting relatives) in an intuitive and safe manner. In the “hospital 4.0” model, the automation system for logistics represents a further integrated node within the ICT network for the management of services, whose architecture is typically distributed and whose management and analysis methodologies are typical of the Data Analysis presented in the previous point. Extending such integration also to a “smart grid” for the management of the utilities, it is possible to increase the energy efficiency of the logistics system. Such an integration also enables a notable containment of the peaks of demanded power to the net (“peak shaving”), with consequent containment of the expense for electric consumption. We want to evaluate the integration of methodologies typical of energy optimization with the management of logistics automation in order to ensure the safe performance of all critical operations by controlling and, if possible, optimizing energy consumption.

1.1.4. Robotics for Health

To ensure continuous and personalized care for patients in the wards or at their homes, solutions involving the use of nurse-robots will be investigated. These intelligent machines will help patients perform simple daily actions, facilitate remote monitoring and communication with

medical staff or relatives, administer simple therapies, or can be used for entertainment (reading, storytelling, playing games) [3].

In addition to nursing robots, devices and control strategies for rehabilitation will be designed such as the development of virtual agents to be implemented in augmented reality that can interact with the patient through advanced techniques of automatic control and provide real-time data to medical staff through telemedicine strategies.

Robotics is already a widespread reality in several medical-surgical specialties. The use of tele-operated or computer-guided machines offers numerous advantages such as precision, repeatability, and tremor filtering [4]. Robots such as the da Vinci system for minimally invasive robotic surgery allow to improve and reduce the duration of the post-operative course of patients. The aim is to improve the capabilities of currently used robots through the use of new sensors, advanced image processing and sensory fusion techniques, computerized procedures for surgery planning based on pre-operative images or guidance through intra-operative image processing, virtual and augmented reality, and new human-robot interfaces [5, 6]. These interfaces, connected to analog or software simulators, will be used, thanks to the already structured collaboration of DIETI within the ICAROS center, for the training of surgeons. New sensorized surgical instruments will be designed and controlled inspired by the human manipulation capability. Anthropomorphic gripping instruments will be developed for both surgery and rehabilitation.

2. The Research Infrastructures

The laboratory is intended as an hub for collecting the computational and storage needs of researchers at Unina. More in details the goal is to provide a complete support in order to profitably leverage the skills of the laboratory team in order to create a room for exchanging ideas and providing multidisciplinary insight on how to build solid benchmarks for assessing research results in several fields. In the following, we will describe the two big data [7, 8, 9, 10, 11, 12, 13] infrastructures that will be available for researchers.

2.1. Public Cloud resources

The eHBDA laboratory in the Cloud is based on virtual infrastructures, i.e. Cloud services, consisting of the following main components: data collection, storage, processing and consumption. Below is a description of the elements composing our infrastructure.

Data collection module. It features 2 Instances of virtual machines each with 8 virtual cores with 3.0 Ghz Xeon Platinum processors providing the possibility to get access to a further increase in performance using Intel Turbo Boost technology equipped with 16 GB of RAM and 20 GB of SSD persistent storage.

Data storage module. It offers a managed Relational Database Service based on MySQL compatible technologies with a minimum space of 2 TB. The service includes the ability to automatically schedule and execute backups and make the necessary tools available for a possible restore; it is also possible to extend the single database instance in order to support Big Data analysis scenarios. The database service is implemented by virtual machines having

the following characteristics: 8 virtual cores with Intel Xeon Ivy Bridge processors with the possibility of having memory bandwidth larger than 60000 MB/s and 122 GB of RAM.

It also provides a No SQL database service providing a storage space of 100 GB, with item size 20KB, data access times less than ten milliseconds, integrated backup and the ability to dynamically scale according to the workload without any service interruption. The Object Storage service guarantees high durability of objects provided in serverless mode and which can scale on demand. The service can also be used as a storage service for the implementation of a datalake. The service includes a storage space of at least 200GB, with the possibility of making at least 2000 Req/month and 2000 Put/month.

Furthermore, there is an in-memory cache service based, on Redis, with available space of at least 60GB. The service is implemented by virtual machines having the following features: 2 virtual cores and 15 GB of RAM.

Finally, a managed service for real-time data ingestion capable of receiving a data flow up to 1MB/s as input is available that allows integration with Spark to allow real-time processing of streaming data by Spark Streaming.

Data processing module. This module provides a managed data warehouse service, with the possibility of scaling (scale-out) the computing cluster in order to manage the increase in the volume of data to be processed. The solution includes back-up functionalities able to manage the DB saving even when the volumes managed grow suddenly. Overall, the computing cluster provide 8 TB of storage space and 8 vcore and 120 GB of RAM. The Business Intelligence service is implemented in IaaS mode by 4 virtual machine instances having the following characteristics each: 4 virtual cores with Xeon Broadwell processors, 30 GB of RAM, 100 GB of SSD persistent storage.

A Managed service for the implementation of a cluster based on Hadoop, Hive, Spark technologies (including the SparkSQL, Spark Mlib, Spark Streaming and GraphX extensions) is integrated with storage solutions for the implementation of the datalake both internal and external, allowing a decoupling of costs and resources dedicated to computing from those dedicated to data storage. The cluster solution is easily integrated with the tools used for the creation of NoSQL databases and Datawarehouse. There are 8 nodes of the Hadoop cluster having the following features: 8 virtual cores with Intel Xeon Ivy Bridge, 60 GB of RAM, 100 GB of storage space in SSD technology, directly attached to the instance plus 100GB of persistent storage.

Finally, a code service allows to write programs in the following languages: Node.js, (JavaScript), Python, Java (Java 8 compatible), and C # (.NET Core), and Go. It checks the limits (throttling) of the functions to prevent programming errors that can generate an unexpected increase in costs and and support the versioning of functions in order to change the definition “at a glance” without impact on the running processes.

Data consume module. The Business Intelligence service in fully managed mode allows the creation of dashboards and that allows integration with the data sources included in the configuration (Datalake, Datawarehouse, Relational Database). The service allows the drafting of fully customizable graphic dashboards, the definition of the analysis paths and the caching of the data in memory to allow a rapid exploration of the same. The service must have at least 10GB of stored data model. Business Intelligence service is based on technologies such as Kibana and Qlikview and supported by at least 2 instances of virtual machines each with the following

configuration: 4 virtual cores with Xeon Platinum 3.0 Ghz processors with the possibility of accessing a further increase in performance using Intel Turbo Boost technology, 8 GB of RAM and 10 GB of SSD persistent storage.

2.2. Private Cloud resources

In order to provide users a broader choice of solutions to fulfill any research needs, we also built our own infrastructure for big data analysis. In the following, we will describe the main features of our infrastructure.

We have 10 compute nodes configured to have a cluster with this features: power supplies for at least 7500W, with at least two hot-swappable power supplies per node, 100 cores in total with 1.80GHz frequency and at least 11M cache per CPU, 1280 GB of total RAM, in banks of at least 32GB DDR4, 4.8 TB total hot plug SSD storage, with drives of at least 480GB, Hot-plug 10TB SAS/SATA storage, with disks of at least 1TB, RAID Controller and Onboard SATA Controller, PCIe x16 expansion slot in each node, Network connectivity of at least 20Gbps for each node and an aggregate (number of interfaces per node x speed per interface x number of nodes) of at least 200Gbps, remote management card on each node with at least 1Gbps dedicated interface, rack mount kit, which must be supplied as specified below.

The storage is guaranteed by 750TB SSD hard disks, having 500.000 IOPS, a latency of 10ms, a CPU with 12 Core, 2.40GHz, 30M Cache, connectivity rate of 4 x10 GbE + 4 x 1 GbE. We have the possibility to scale-up to 1PB with an availability of 99.9999% by a dedicated switch having 16 10GbE port, 2 25GbE/100GbE port.

In order to obtain a high performance level, we built a proper network infrastructure as follows. A Switch for the production network, including 3mt cables for connection to the nodes, with the following characteristics: 48 25GbE SFP28 ports + 6 100GbE ports, 3.6 Tbps (full-duplex) non-blocking, store and forward switching fabric, L2 and L3 Ethernet switching with support for QoS, features for IPv4 and IPv6, including support for OSPF and BGP routing, Support for OpenFlow v 1.3. A Switch for the management network with 48 1GbE Base-T ports + 4 10GbE ports and 3mt RJ45/RJ45 cables for connection to the nodes. Finally, we have 3 Rack 750x1200mm able to contain all the nodes described above (computing nodes, storage nodes and switches).

3. Conclusion

In this paper, we described the e-health laboratory we implemented at DIETI in order to support a variety of big data analyses for a broad set of application scenarios [14]. The goal of this paper is to provide a quick view of an actual infrastructure that could be used as a reference architecture for top-class applications.

Acknowledgments

Supported by the Ministry of University and Research for Department of Excellence Project.

References

[1] G. Aceto, A. Botta, A. Pescapé, C. Westphal, Efficient storage and processing of high-volume network monitoring data, IEEE Transactions on Network and Service Management 19 (2013) 1–14.

[2] K. Wang, Y. Shao, L. Shu, C. Zhu, Y. Zhang, Mobile big data fault-tolerant processing for ehealth networks, IEEE Network 30 (2016) 36–42.

[3] S. Chiaverini, B. Siciliano, L. Villani, A survey of robot interaction control schemes with experimental comparison, IEEE/ASME Transactions on Mechatronics 4 (1999) 273–285. doi:10.1109/3516.789685.

[4] A. Mashayekhi, S. Behbahani, F. Ficuciello, B. Siciliano, Influence of human operator on stability of haptic rendering: a closed-form equation, International Journal of Intelligent Robotics and Applications 4 (2020). doi:10.1007/s41315-020-00131-6.

[5] A. Botta, L. Gallo, G. Ventre, Cloud, fog, and dew robotics: Architectures for next generation applications, in: 2019 7th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), IEEE, 2019, pp. 16–23.

[6] G. Stanco, A. Botta, G. Ventre, Dewros: a platform for informed dew robotics in ros, in: 2020 8th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), IEEE, 2020, pp. 9–16.

[7] V. Persico, A. Pescapé, A. Picariello, G. Sperlí, Benchmarking big data architectures for social networks data processing using public cloud platforms, Future Generation Computer Systems 89 (2018) 98 – 109. URL: http://www.sciencedirect.com/science/article/pii/S0167739X17328303. doi:https://doi.org/10.1016/j.future.2018.05.068.

[8] D. Agrawal et al., Challenges and opportunities with big data. A community white paper developed by leading researchers across the United States (2012).

[9] V. R. Borkar, M. J. Carey, C. Li, Inside "Big Data Management": Ogres, Onions, or Parfaits?, in: International Conference on Extending Database Technology, 2012, pp. 3–14.

[10] T. Economist, Data, data everywhere, The Economist (2010).

[11] V. L. Heron, Michaeland Hanson, I. Ricketts, Open source and accessibility: advantages and limitations, Journal of Interaction Science 1 (2013) 1–10. URL: http://dx.doi.org/10.1186/2194-0827-1-2. doi:10.1186/2194-0827-1-2.

[12] Nature, Big data, Nature (2008).

[13] S. Mgudlwa, T. Iyamu, Integration of social media with healthcare big data for improved service delivery, SA Journal of Information Management 20 (2018). doi:10.4102/sajim.v20i1.894.

[14] G. Manco, E. Masciari, A. Tagarelli, A framework for adaptive mail classification, in: 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2002), 4-6 November 2002, Washington, DC, USA, IEEE Computer Society, 2002, p. 387. URL: https://doi.org/10.1109/TAI.2002.1180829. doi:10.1109/TAI.2002.1180829.