Advancing Integrated and Personalized Healthcare Services, the AEGLE Approach

Andreas Raptopoulos
Exodus AE
arap@exus.co.uk

Vassilios Tsoutsouras and Dimitrios Soudris
School of Electrical and Computer Engineering, National Tech. Univ. of Athens, Greece
{billtsou, dsoudris}@microlab.ntua.gr

***Abstract*— The AEGLE project aims to advance integrated and personalized healthcare services, by innovatively handling big-biodata both at the cloud and at local healthcare sites. At the local level, AEGLE will focus on real-time processing of large volumes of raw data originating from patient monitoring services. Then at the cloud level, AEGLE will offer an experimental big data research platform to data scientists, workers and data professionals across Europe. This paper presents the AEGLE's approach to healthcare, along with the medical test cases and underlying technologies used in the project.**

Keywords—healthcare, bigdata, dataflow acceleration, cloud computing

I. INTRODUCTION

Nowadays, there is an obvious gap in the area of big data analytics for Health Bio-data. Data-driven services are still needed to cater for the data versatility, volume, velocity and veracity within the whole data value chain of healthcare analytics. A true opportunity exists to produce value out of big data in healthcare with the goal to revolutionize integrated and personalized healthcare services.

Modeling biological phenomena is typically very complex and has always been understood to be a computationally intensive process. In order to draw meaning from the exponentially increasing quantity of healthcare data, it must be dealt with from a big data perspective, using technologies capable of processing massive amounts of data efficiently and securely. Collecting and aggregating anonymous data from geographically dispersed locations makes it possible to construct statistically meaningful databases, based on which macroscopic reasoning can be made, rather than solely focusing on the individual and associated pathology.

Several European initiatives [1] have already pinpointed the importance and usefulness of healthcare big data, e.g. to predict the outbreak of an epidemic etc. Additionally, business interest is growing like the Open Data initiative, where health big data providers, governmental and research institutes and industry aim to develop a vendor-neutral Big-Data platform [2]. Organizations are taking a serious view on big data, recognizing the critical success factors and issues associated with handling enormous volumes of data. Big data not only is a major challenge for ICT and healthcare professionals, but also is a great societal opportunity. The use of massively available medical data may allow clinicians to simulate potential outcomes and so prevent patients from undergoing ineffective treatments or make them better treated. In other words, accumulating and using data to develop a greater understanding of pathophysiological processes will result in significant healthcare improvements.

The AEGLE project [3] targets to address the aforementioned open issues by implementing a full data value chain to create new value out of rich, multi-diverse, big health data. AEGLE's mission is to realize a European business ecosystem to healthcare stakeholders, industry and researchers for creating out-of-box knowledge in order to provide cloud and HPC data services and support new products that will improve health. The project builds upon the synergy of heterogeneous High Performance Computing (HPC), Cloud and Big Data computing technologies for the delivering optimized analytic services on Big-Bio Data application use cases from the medical and health-care domain.

II. AEGLE IN REPSECT TO OTHER R&D PROJECTS ON HEALTHCARE

Currently numerous R&D projects are running, regarding health and ICT technologies. Most of them are targeting to obtain a proof of concept on the impact of sensing and monitoring devices in the treatment and management of a disease. Some of the projects have already examining in more depth the concept of integrated care concerning chronic diseases like WELCOME [4] or SWAN-iCare [5]. Additionally health projects have started exploiting cloud capabilities, like e-health GATEway to the Clouds [6] or BIOBANK Cloud [7] as well as perform large scale analysis like MD-Paedigree [8] and OPENPHACTS [9]. Most of these projects however aim to the storage and analysis of mainly biological data (e.g. genomics), and this is the field where commercial products can be found like CLCbio [10] platform aiming to the analysis of DNA, RNA and protein.

Fig. 1 illustrates the positioning of AEGLE project in respect to other projects on eHealth and Big Data for health-care. As it can be seen, none of the existing Big-Data projects are completely dedicated to healthcare and the provision of corresponding healthcare services, or the management of diseases. AEGLE combines all elements of the full value chain (storage of large volumes of data, big data analytics, cloud computing and provisioning of integrated care services), targeting to cover the whole field of health big data analytics. It will also liaise with other projects (e.g OPENPHACTS etc), for

taking advantage from their developments, resulting to a more advanced and extended system.

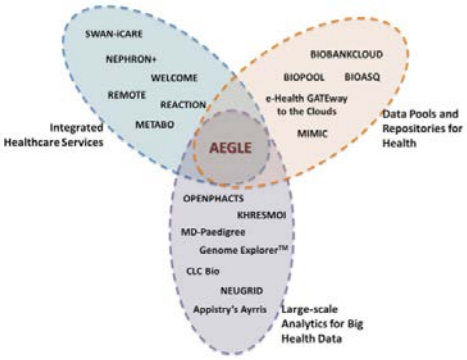

Fig. 1. AEGLE's position compared to other related projects

III. The Big Bio-data Challenge as seen from AEGLE's perspective

Big data in health refers to electronic health-related data sets that cannot be managed with traditional software and/or hardware and common methods. Big data is bringing challenges to traditional data processing, as regards the size of data (volume), the required processing speed (Velocity), the heterogeneity (variety) and the accuracy (veracity).

The most crucial challenge for the success of Big Data in Health is to make value out of these data. Health research has been built on small and clean data, with carefully designed cleaned trials and extrapolation of their findings. A shift from hypothesis driven to data driven research is foreseen, based on machine learning techniques that mine patterns, clusters and associations for big (e.g. population representative) volumes of unclean data. To increase medical credibility, the produced knowledge and hypotheses can then be confirmed in smaller and cleaner datasets. The medical cases of AEGLE have been carefully chosen to cover biomedical research and questions that can set the basis for bio-signal and bioinformatics analytics, multiparametric pattern mining, and integrative predictive modelling. Presentation of information and visualization techniques for a multitude of medical data types, and their interconnections, will be another complexity level. Finally, in AEGLE the path from data to knowledge, interpretation, actionable data, necessarily involves open data standards for sharing and interoperability, methods for semantic and temporal similarity, as well as standardized integration of data, e.g. clinical and genomic. Applying, interlinking and extending current medical standards for the integrated and quality based use or even repurposing of big biodata will be a key issue in AEGLE, addressing both variety and veracity.

A. *Medical cases considered in AEGLE*

Three medical cases are considered in the AEGLE project as representative cases for dealing with big-data in healthcare.

Chronic Lymphocytic Leukaemia (CLL): CLL is a chronic, incurable disease, leading to great distress for patients and their families as well as huge costs for the health care system. Analysis will be performed to address complex clinical questions and scenarios associating phenotypic data with personal genetic profiles. In addition, AEGLE will offer the possibility of proposing and evaluating health interventions towards the goal of integrated care e.g. identifying groups with specific profiles that will be considered as eligible or ineligible for certain treatments and, at the same time, evaluating the cost of this intervention.

Intensive Care Unit (ICU): In an Intensive Care Unit context, patient bio-signals are continuously monitored and displayed towards recognizing alerting events. The recordings of clinical, laboratory data and physiologic waveforms could be analysed and displayed in an easy-to-understand manner for clinicians. AEGLE's scalable data analytics will provide automated analysis of the fast changing multi-dimensional functions of variables for the detection of unusual, unstable or deteriorating states in patients. In this respect, early and personalized treatment will be feasible using AEGLE technology for higher survival in ICUs around European Hospitals.

Type 2 Diabetes (T2D): The risk of developing T2D can be increased by various factors; usually a mixture of modifiable and non-modifiable elements of age, weight, genetics and ethnicity. The AEGLE system will analyse the inter dependences of the factors including medication that are known to have a detrimental effect in type 2 diabetes to give a prediction on the potential deterioration. This would enable intervention to enable reduction of mortality, complications and hospitalization that would all lead to reduction in overall health costs.

B. *AEGLE's system architecture*

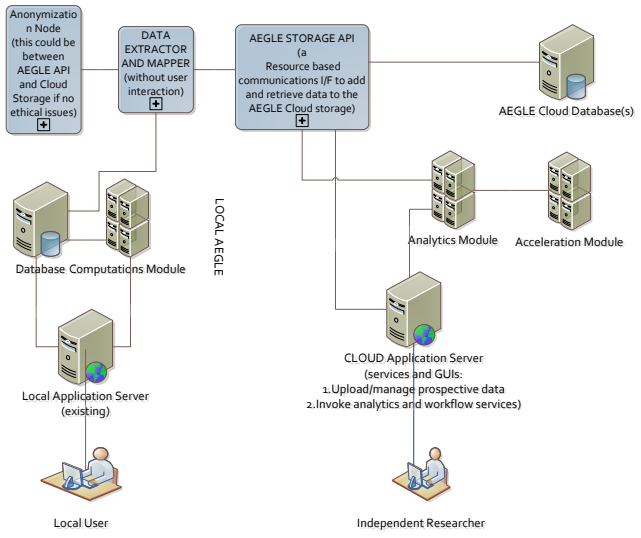

Fig. 2. System architecture envisioned in AEGLE

Fig. 2 depicts the main building blocks of AEGLEs big data analytics framework. Reflecting the requirements of different stakeholders involved in the full data value chain for healthcare analytics, the AEGLE framework consists of big data analytics services at two levels.

Local level: The local level implements big data analytics services for real-time processing of large volumes, fast generated and multiple-formatted raw data originating from patient monitoring services deployed within a healthcare unit, complemented with dedicated medical databases. An example is the real-time analytics service that AEGLE will implement for the scenario of ICU. The goal of the analytics service is to detect unusual, unstable or deteriorating states of patients given the fast changing multi-dimensional variables conveyed within the bio-signals generated by ICU dedicated equipment. The stakeholders of the local level analytics are healthcare units/systems and of course the patients as the ultimate beneficiaries that will benefit from the advanced treatment modalities enabled by adopting the analytics services. For example, in the ICU scenario, a prompt reaction to detected instabilities or abnormal behavior of the patient's status could significantly help to save the lives of patients being treated within the ICU.

Cloud level: The cloud level analytics services will offer an experimental big data research platform to data scientists, workers and data professionals across Europe. The platform consists of a large pool of semantically-annotated and anonymized healthcare data, a set of libraries implementing state-of-the-art big data analytics methods including the local level big data analytics AEGLE services and APIs for federating with public and private data sets. Advanced visualization tools will be implemented by AEGLE as an instrument for gaining new knowledge and expertise, advancing the European know-how in healthcare big data analytics, by allowing data scientists to steer the cloud level analytics mechanisms with their own insights. SMEs across Europe will be given the ability to use the AEGLE platform in order to deploy and assess the validity of their innovative data analytics solutions.

IV. Underlying technologies to Support AEGLE's Vision

A. Big-data frameworks

Hadoop Framework: The Hadoop Distributed File System (HDFS) [11] has been developed by Apache, as part of the Apache Hadoop Core project [12]. Applications that run on HDFS have large datasets, meaning that a typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files. It provides high aggregate data bandwidth and scales to hundreds of nodes in a single cluster. It can potentially support tens of millions of files in a single cluster instance. HDFS has been designed to be easily portable from one platform to another. This facilitates widespread adoption of HDFS as a platform of choice for a large set of applications.

MapReduce Programming Paradigm: The MapReduce paradigm [13] has emerged as a popular approach to handling large-scale analysis, farming out requests to a cluster of nodes that first perform filtering and transformation of the data (map) and then aggregate the results (reduce). MapReduce is a software framework for easily writing applications which process vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault tolerant manner. The MapReduce framework that has been implemented by Apache, is designed to run on top of an HDFS cluster deployment. The datasets that are processed by MapReduce jobs can potentially scale to several terabytes. MapReduce jobs can utilize clusters that consist of hundreds or even thousands of nodes.

A MapReduce job usually splits the input data set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. The map tasks process key/value pairs to generate a set of intermediate key/value pairs and the reduce tasks merge all intermediate values associated with the same intermediate key, so as to produce the final key value pairs, which are the output of the MapReduce job. The input and the output of a MapReduce job are stored in HDFS. This decision allows the framework to effectively schedule tasks on the nodes where the data is already present, resulting in very high aggregate bandwidth across the cluster. The MapReduce framework takes care of the details of partitioning of the input data, scheduling the programs execution across a set of machines, handling machine failures and handling the required inter machine communication. Thus, it provides a level of abstraction that hides the messy details of parallelization, fault tolerance, data distribution and load balancing, letting programmers to express the simple computations that they are trying to perform.

B. Medical data anonymization

As a new generation of big data healthcare platforms, AEGLE promotes a novel approach to extraction, desensitization and sharing of medical data in a collaborative manner. To do so AEGLE implements at its core the principle of "Privacy by Design" and the use of aggregate data, instead of raw data, thus ensuring the highest level of confidentiality. AEGLE also leverages on legacy assets from its key partners, each bringing key technological components in the final solution.

Amongst these and developed in collaboration with renowned medical centers in Europe, FedEHR from GNUBILA [14], is a patient-centric Electronic Health Records (EHR) big data solution, supporting this long-term goal. FedEHR, stands for Federated EHR. It leverages on the cloud elasticity to provide a scalable vendor-neutral anonymization database able to cope with massive multi-modal and heterogeneous medical information, data and knowledge integration. FedEHR takes its roots in leading edge technologies developed and tested in computationally and data intensive environments at the European Organization for Nuclear Research (CERN) [15].

At its very core, is an innovative anonymization machinery, which couples 4 major data mining techniques to identify personal information and treat it accordingly. FedEHR anonymizer is thus able to deeply scrutinize different types of data and formats, in order to spot sensitive information areas from metadata to data, and to alter them [16]. Thanks to anonymization profiles, which data curators can define based on ethical concerns and applicable regulations, FedEHR automatically treats targets by replacing, removing, modifying or encrypting information. FedEHR combines regular data processing with approximate search and natural language processing to achieve an in depth anonymization, towards the creation of an anonymous and homomorphous representation

of the patient, which can then be shared and processed in large-scale cross-enterprise and transnational studies.

C. Dataflow computing for big-data acceleration

Rapid developments in data-collection technologies, storage capabilities and networked digital devices have lead to the advent of very large data sets that are difficult to process and analyze with conventional approaches. Conventional scaling approaches of simply adding more processing nodes to the data center can reach their limitations in available space, and power efficiency is also becoming increasingly import in terms of both cost and environmental impact of computing.

One solution to achieve faster and more efficient Big Data Analytics processing lies in using a new computing paradigm. Instead of writing a program that describes a sequence of instructions on data we can write a program that describes the flow of data through a structure of highly optimized operators: i.e. dataflow computing. This computing model is similar to an assembly line on a factory floor where parts (data) arrive just in time at dedicated workstations (arithmetic operators) and move forward in lockstep to produce final products (results). Compared to a conventional von-Neumann processor, this model of computation is much more efficient for many large-scale applications since the movement of data is minimized, and auxiliary components such as instruction decoding logic, branch prediction units, and general purpose caches are eliminated. Maxeler Technologies [17] commercializes this approach in its high-performance multi-scale dataflow computing technology. Maxeler's multidisciplinary approach to high-performance, high-efficiency computing enables a team of domain experts such as scientists, analysts or engineers to formulate and optimize their algorithms in a high-level dataflow oriented language. Targeting a Maxeler dataflow computer typically results in 20-50× improvements in terms of both performance and power efficiency over conventional server technology with the same physical dimensions.

AEGLE will utilize Maxelers dataflow computing for fast and efficient Big Data Analytics processing. Dataflow acceleration will be applied to three different levels. At the algorithmic level, customized dataflow engines (DFEs) will be explored [18] and developed to accelerate the compute intensive kernels found in the targeted Big Data Analytics procedures. At the runtime level, specialized DFEs will be designed targeting the acceleration of the underlying MapReduce programming model, i.e. the map, combine and reduce functions. In addition, customized memory management schemes [19] will be incorporated to efficiently handle the large number of key-value pairs usually generated by MapReduce semantics, as well as platform specific task schedulers for balancing the load across the software processors and the DFEs. Finally, at the storage and data management level, the database management system (DBMS) will be extended to support both adaptive data layout optimizations as well as query-specific dataflow-based acceleration of compute intensive database operations. The integration of the dataflow accelerated Big Data services will be incorporated in a transparent manner in the final AEGLE system architecture. Efficient dataflow acceleration will bring the benefit of improved processing speeds, reduced area and power as well as lower cost than standard server systems.

V. Acknowledgements

This work partially supported by the H2020 project "Aegle", http://www.aegle-uhealth.eu and DAAD-funded project "Teacher".

VI. Conclusions

In this paper, we presented the AEGLE approach for advancing integrated and personalized healthcare, while enabling high performance Big Bio-Data analytics. AEGLE aims to efficiently integrate cloud computing together with heterogeneous high performance computing technologies to enable both a publicly available global medical repository for wide adoption within the healthcare research, as well as to support fast analysis for aiding medical decisions at the local level of intervention. Three advanced Big Data use cases were presented along with the key enabling technologies to support AEGLE's vision.

References

[1] Big data: What is it and why is it important? Available online http://ec.europa.eu/digital-agenda/en/news/big-data-what-it-and-why-it-important.

[2] Top Big Data opportunities for health startups. http://www.healthstartup.eu/blog/top-big-data-opportunities-for-health-startups/.

[3] AEGLE: An analytics framework for integrated and personalized health-care services in Europe, http://www.aegle-uhealth.eu

[4] WELCOME EU project, http://www.welcome-project.eu/

[5] SWAN-iCare project, http://www.swan-icare.eu

[6] GATEway to the clouds, http://www.jisc.ac.uk/whatwedo/programmes/diresearch/researchtools/ehealth.aspx

[7] BIBANK Cloud, http://www.biobankcloud.com

[8] MD_Paedigree, http://www.mdpaedigree.eu

[9] OPENPHACTS, http://www.openphacts.org

[10] CLCbio, http://www.clcbio.com

[11] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[12] APACHE, Hadoop framework, https://hadoop.apache.org

[13] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[14] GNUBILA, www.gnubila.fr

[15] D. Manset. From physics to daily life, application in biology, medicine, and healthcare, chapter 11, p. 233, wiley, 2014. cern 60th anniversary.

[16] Berlanga R., Manset D, et al. Medical data integration and the semantic annotation of medical protocols. 21st ieee international symposium on computer-based medical systems (cbms 2008). university of jyvskyl, finland, june 17-19, 2008. springer-verlag isbn 3-540-48273-3

[17] MAXELER Technologies, http://www.maxeler.com

[18] Sotirios Xydis, Kiamal Pekmestzi, Dimitrios Soudris, and George Economakos. Compiler-in-the-loop exploration during datapath synthesis for higher quality delay-area trade-offs. *ACM Trans. Des. Autom. Electron. Syst.*, 18(1):11:1–11:35, January 2013.

[19] S.Xydis,A.Bartzas,I.Anagnostopoulos,D.Soudris,andK.Pekmestzi. Custom multi-threaded dynamic memory management for multiprocessor system-on-chip platforms. In *Embedded Computer Systems (SAMOS), 2010 International Conference on*, pages 102–109, July 2010