AEGLE: A Big Bio-Data Analytics Framework for Integrated Health-Care Services

Dimitrios Soudris[1], Sotirios Xydis[1], Christos Baloukas[1], Anastasia Hadzidimitriou[2], Ioanna Chouvarda[2], Kostas Stamatopoulos[2], Nicos Maglaveras[2], John Chang[3], Andreas Raptopoulos[4], David Manset[5], Barbara Pierscionek[6], Reem Kayyali[6], Nada Phillip[6], Tobias Becker[7], Katerina Vaporidi[8], Eumorphia Kondili[8], Dimitrios Georgopoulos[8], Lesley-Ann Sutton[9], Richard Rosenquist [9], Lydia Scarfo[10], Paolo Ghia[10]

[1]Institute of Communications and Computer Systems, GR,
[2] Centre for Research and Technology Hellas, GR,
[3] Croydon Health Services National Health Service Trust, UK,
[4] EXODUS S.A, GR,
[5] GNUBILA, FR,
[6] Kingston University Higher Education Corporation, UK,
[7] MAXELER Technologies, UK,
[8] University Hospital of Heraklio, GR,
[9] Uppsala University, SW,
[10] University Vita-Salute San Raffaele

***Abstract*—AEGLE project[1] targets to build an innovative ICT solution addressing the whole data value chain for health based on: cloud computing enabling dynamic resource allocation, HPC infrastructures for computational acceleration and advanced visualization techniques. In this paper, we provide an analysis of the addressed Big Data health scenarios and we describe the key enabling technologies, as well as data privacy and regulatory issues to be integrated into AEGLE's ecosystem, enabling advanced health-care analytic services, while also promoting related research activities.**

I. Introduction

At the centre of health debates there are open questions on how to manipulate data and how to produce value out of it, share it and secure it [1]. Although, the term Big-Data has become a buzzword in the field of information technology, its applicability on biological and health-based data, that naturally quite complicated and difficult to collect, is still limited. 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 [2] 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 [3]. 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. However, the strategic advantage brought by Big-Data in healthcare still materializes at slow paces, as only some large-scale organizations have established few pilot or proof-of-concept projects.

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 personalised healthcare services. The AEGLE project 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 an 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. In this paper, we describe in depth the three target Big-Bio

[1]This research is partially supported by the E.C. funded program AEGLE under H2020 Grant Agreement No: 644906, http://www.aegle-uhealth.eu

Data applications as well as the key technologies to be utilized within AEGLE for delivering accelerated health-care analytics.

The organization of the paper adopts the following structure: Section II refers to current state-of-art showing positioning of AEGLE project in respect to other funded research projects in the fields of health-care and Big-Data. In Section III, we discuss the main features of the Big Data domain and show how these features maps to the requirements of modern large scale health-care services. Section V analyzes in depth the AEGLE's use cases, i.e. modern medical applications stretching the adoption of advanced analytics services. AEGLE's architecture is presented in Section IV, while Section VI describing the main components and technologies utilized. Section VIII concludes the paper.

II. AEGLE's relevance and positioning in respect to other R&D projects

AEGLE aims to generate value from healthcare data with the vision to improve translational medicine and facilitate personalized and integrated care services overall improving healthcare at all levels, to promote data-driven research across Europe and to serve as an enabler technology platform. It will enable business growth in the field of big data analytics for healthcare Currently numerous R&D projects are running, regarding health and ICT technologies. Most of them are targeting to obtain a proof of concept (having limited and controlled validation phases) on the impact of sensing and monitoring devices in the treatment and management of a disease.

Some of the projects have already examined in more depth the concept of integrated care concerning chronic diseases like WELCOME[2] or SWAN-iCare[3]. Additionally health projects have started exploiting cloud capabilities, like e-health GATEway to the Clouds[4] or BIOBANK Cloud[5] as well as perform large scale analysis like MD-Paedigree[6] and OPENPHACTS[7]. 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[8] platform aiming to the analysis of DNA, RNA and protein.

Figure 1 illustrates the positioning of AEGLE project in respect to other projects on eHealth and Big Data for healthcare. 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 in a more advanced and extended system.

[2]http://www.welcome-project.eu/

[3]http://www.swan-icare.eu

[4]http://www.jisc.ac.uk/whatwedo/programmes/di_research/researchtools/ehealth.aspx

[5]http://www.biobankcloud.com

[6]http://www.md-paedigree.eu

[7]http://www.openphacts.org

[8]http://www.clcbio.com

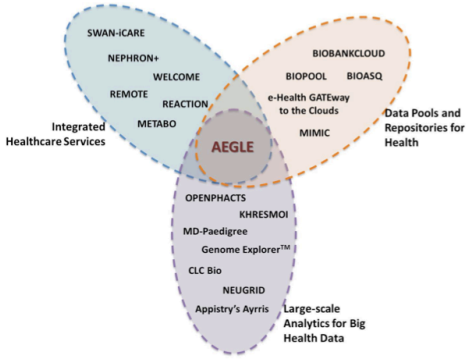

Fig. 1. Positioning in respect to other R&D projects on healthcare.

III. Big Data for Health and the AEGLE 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).

Indeed health data volume is expected to grow dramatically and even now in its totality is overwhelming because of its volume. These include different types and analytics scope: a) EHR longitudinal data useful for predictive analysis and deep phenotyping, along with multimedia clinical data medical imaging, laboratory, pharmacy, insurance, b) personal quantifiable and social data in telehealth, social media and quantified self domain, c) -omics data, d) medical knowledge and literature.

Although ICT technology makes advancements to give solutions in big data volume and velocity issues, the healthcare industry has been hesitant in embracing Big Data. Privacy, openness and sharing concerns are raised in health related data. Their expected heterogeneity and poor quality has to be overcome for veracity, while paradigm shifts towards data-driven analytics have to be accepted. AEGLE aims to address these bottlenecks and contribute towards filling these gaps in a range of medical problems, that will be described in detail in the next Section.

A typical distinction in BD management concerns their characterisation as structured that follow an explicit data model (like the EHR, biosignals) and unstructured data (text and images), which may not have a clear border, but emphasizes in the ease of interpretability. Clinical data from medical records, streaming biosignals and omics data will be collected in Aegle, covering a big volume and a wide range of data types in the three medical cases, and addressing problems that require integrative analytics. While most data will be structured, challenges will arise in text narratives of health records, medical images, as well as the multitude of semi-structured omics data. Apparently, besides volume, the complexity and diversity of data types that can be derived from these sources is a challenge, while analytics performance (velocity) will be required due to not only the real time nature of clinical questions (e.g. critical care alerting) but also due to the need

to support in a realistic manner research computational needs (e.g. genomic analysis).

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 three medical cases of AEGLE have been carefully chosen to cover biomedical research and questions that can set the basis for biosignal and bioinformatics analytics, multiparametric pattern mining, and integrative predictive modelling. Presentation of information and visualisation 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.

IV. The AEGLE System Architecture

Figure 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 Intensive Care Unit (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 are 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 behaviour of the patients 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 visualisation 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 (according to the business model that AEGLE will define) in order to deploy and assess the validity of their innovative data analytics solutions which aim at creating new value in the field of healthcare.

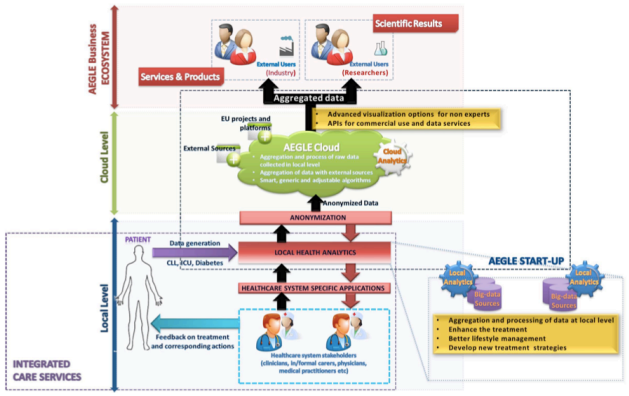

Fig. 2. AEGLE infrastructure.

V. AEGLE Use Cases

1) CLL description: Chronic lymphocytic leukemia (CLL) is a chronic malignancy of B lymphocytes displaying remarkable biological and clinical heterogeneity. CLL is more frequent in the West [Europe [4] and North America], and much more rare in the Far East. Low incidence rates for CLL persist in individuals migrating to the U.S. from Asian countries and in their descendants, suggesting a strong genetic component, also supported by the well-established inherited familial predisposition to CLL [5]. The incidence of CLL in men is twice that reported for women; the underlying cause for this gender bias remains elusive. The median age at diagnosis is 72 years, however up to 20-30% of patients referred to specialized hematological centers, are aged 55 or younger [5].

2) Analysis of data characteristics: In recent years, advances in CLL research have offered a wider picture of the disease, encompassing its links to monoclonal B lymphocytosis (MBL), a very frequent, potentially pre-leukemic condition [6]. CLL is now considered a prototype for cancers in which tumor initiation and progression are linked to the interplay between cell-intrinsic and cell-extrinsic mechanisms. The former include more general cancer-associated mechanisms e.g. activation of oncogenes and/or elimination of tumor suppressor genes [7]. The latter relate to the crosstalk of CLL cells with the surrounding microenvironment through various receptors, which mediate signal transmission from the external milieu to the cell interior [8].

Due to the fact that CLL predominantly affects the elderly population and that for many patients the disease remains indolent, CLL management has largely adopted an active monitoring approach i.e. wait and watch. On the contrary, a

portion of patients will become in need for therapy according to established and widely used clinical criteria [9]. For them, several treatment options are available with varying efficacies (from palliation or mere amelioration of symptoms to more effective long-term disease control) and varying costs (from hundreds to tens of thousands of euros). Regarding the latter, the annual costs of providing care to patients with CLL have been estimated as almost double compared to those of matched controls without cancer [10].

3) Limitations of existing approaches suggesting Big Data Analytics: The recent advent of drugs targeting CLL pathophysiological pathways rather than generic cancer-associated mechanisms represents a major paradigm shift having the potential to radically change CLL natural history [11]. However, many issues remain regarding the comprehensive appreciation of their pleiotropic modes of action, their optimal use, long-term efficacy and, no less significant, their very high cost. As the vast majority (up to 80-85%) of all CLL patients are asymptomatic at diagnosis, with the disease being suspected by a routine blood test revealing an increase in lymphocytes, the need exists to achieve an accurate clinical decision and define at the time of the first diagnosis the long-term outcome of each individual patient. This prompts the quest for markers/features of prognostic and predictive value. Many such clinical and biological markers have been identified and found to carry an independent prognostic value, however, robust and personalized prognostication in CLL remains elusive [12]. In particular, issues exist regarding (i) the discordance between markers; (ii) their standardization; (iii) their temporal evolution and links to CLL progression; and (iv) their clinical validation.

Focusing on biomarkers, with the advent of high-throughput technologies, a wealth of information has become available about CLL at unprecedented scale in medical research [13]. The trade-off, being a serious bottleneck, concerns the ability to manage big biodata and other associated information in meaningful ways, while it is growing in size and complexity at phenomenal speeds.

AEGLE aims to overcome these limitations through the development of integrative models to discover and explore links between multiple layers and scales of information, eventually assigning meaning to big biodata. This will enable interrogation, via multi-variant analysis of causes for the differential evolution patterns observed in this disease, which is highly relevant given that CLL is both clinically heterogeneous and incurable. Particularly for MBL, AEGLE aims to create a model that will clearly distinguish those cases that are more likely to progress to CLL and require special clinical attention during follow-up, while sparing most cases from psychological distress and continuous medical attention. This has obvious social and economic implications, ensuring individuals with MBL that their condition is managed appropriately, also affecting and/or guiding lifestyle decisions and providing the authorities and insurance companies with accurate tools for implementing effective policies.

A. Big Multi-parametric Management for Type 2 Diabetes

1) Type 2 Diabetes description: Type 2 Diabetes (T2D) is a chronic condition that arises when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin produced. Diabetes is the fourth leading cause of death in most developed countries. In 2003, the International Diabetes Federation[9] estimated that approximately 194 million people around the world had diabetes. By 2025 this figure is expected to rise to 333 million, amounting to 6.3% of the worlds population living with diabetes. The risk of developing T2D can be increased by various factors; usually a mixture of modifiable and non-modifiable elements of:

1) Age: risk increases with age however people aged 40 years and above are at a higher risk [14].
2) Weight: the larger the weight/waist circumference, the greater the risk [15]. According to the UK National Institute for Health and Clinical Excellence (NICE) guidelines, adults with a Body Mass Index (BMI) of 25kg/m2 or greater are at a higher risk of developing T2D [16].
3) Genetics: the closer the family connection, the greater the risk.
4) Ethnicity: people with certain ethnicities e.g. South Asian, Black-African, African-Caribbean, have a higher risk of developing T2D and subsequently the complications it may bring.

It is important to note that T2D and the associated condition of hyperglycemia is one the most important modifiable factors increasing a persons risk of future cardiovascular disease (CVD) [17]. Furthermore, a linear relationship exists between mean systolic blood pressure (SBP) and the risk of macro- and microvascular complications for diabetes [18]. There are also a number of other biomarkers such as cholesterol, cytokines (interlukin-6) and C-reactive protein. In addition, T2D can lead to complications in the eye (retinopathy), kidneys (nephropathy) and blood vessels in the extremities (peripheral arteriopathy). Treatment of T2D is not always predictable as patients vary in their response to medications and in adverse reactions. There is therefore a continual need to monitor, alter and refine what a particular patient is taking to ensure the maximum benefit for treatment while reducing or ultimately eliminating the risks of adverse reactions and exacerbations. Patient compliance also needs to be watched as this can be the cause of disease progression.

Hence there is a need to develop of an effective screening, monitoring and disease modification strategy for T2D that takes into account all known risk factors such as age, ethnicity, socio-economic and demographic characteristics as well as lifestyle factors and the changing biophysiological profile of the patient with T2D and the particular course of their illness. The use of Big Data analysis should elucidate the major determinants and potential interactions that lead to disease progression, enabling preventative strategies to be established for preservation of health and wellbeing, and to inform appropriate and effective lifestyle interventions to prevention progression of T2D and its complications.

In AEGLE project we consider two UK T2D medical databases that separately and collectively provide a large number of cases with a full gamut of medical and associated data for each case. They are regularly updated. The databases include structured and unstructured factors like:

1) Age, Gender, Medications

[9]http://www.idf.org

2) Anthropometric markers: height, weight
3) Clinical measurements: blood pressure
4) Biochemical markers : plasma glucose, HbA1c, electrolytes, lipids, liver enzymes, urine microalbumin
5) Social and demographic factors : education, area of residence
6) Lifestyle factors: smoking, alcohol consumption, physical activity
7) Clinical notes, images e.g fundus images (in a small subgroup with complications), Doppler sonography and lab tests, e.g. creatinine/albumin ratio, total cholesterol

2) Analysis of data characteristics: Based on the risk factors and screening measurements outlined, and their Variety from biochemical data, ultrasound, to general demographic data it is necessary to develop a scalable infrastructure such as the AEGLE platform that will process streams of distributed and heterogeneous biomedical, imaging and demographic Big Data, while offering multi-parametric data analytics services. Such an infrastructure will cope with important computational issues generated by: a) the complex and heterogeneous, huge volumes of data and, b) the inherent complexity of the multi-parametric data analyses to provide meaningful data to multi-level of healthcare professionals, researchers to inform diabetes early diagnoses, screening and counselling. It must be noted that high-throughput data (Velocity) are usually provided in semi-structured or unstructured format and hence the AEGLE platform will need to provide a multi-step processing modification and analysis of data in order to be operational for the end users. The processing of all the above prognostic and complication indicators in their variety should enable the early detection, treatment and modifications of complications developing in T2D, leading to a decrease in morbidity, mortality and excessive health care costs.

3) Limitations of existing approaches suggesting Big Data Analytics: Currently there are few databases of diabetic patients, as many of these details are contained within larger medical databases. Those that exist do not contain the same details and do not communicate with other databases to enable a cohesive and complete multivariate analysis to be made. It is not currently possible to integrate data easily across databases and to include information from latest advances in medical science enabling assessments of predictive factors such as genetic predisposition. Moreover, one of the main challenges in big data management and analysis is the Veracity of data in terms of data integrity and incomplete data sets in existing datasets.

The AEGLE system will analyse the inter dependences of the variables 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 which would all lead to reduction in overall health costs. The process will take place mainly in AEGLE cloud, and corresponding visualization options will be provided.

Big data analytics in AEGLE will enable statistical analysis of possible interactions of co-morbidities in type 2 diabetes, via multi-variant analysis, such that small but statistically and clinical relevant significance can be identified to enable exploration of new modalities of intervention that may modify disease complication. Also, it may be possible, as in genetic studies to date, to be able to identify novel genetic mutations that may predispose to pharmacologic interaction, that may prevent or reduce the rate of disease progression. By using a big data, it is possible to identify any new implemented changes quickly to enable decisions to be made early on in regards to the usefulness of the intervention, thus leading to early discontinuation of non-viable products in favour of promising new developments.

In summary the use case of T2D in AEGLE is to give deep insight into disease progression, to enable researcher and clinicians to make better decisions about patient health. It will use big data analytics to predict outcomes of medical complications for individuals that really impact the lives of people living with diabetes heart attacks, strokes, eye disease, kidney disease and limb amputations. The ultimate outcome sought is to predict and improve on cost of care, hospital readmission and mortality.

B. Intensive Care Unit (ICU) BIG Bio-signal streams and Challenges

1) ICU description: In an Intensive Care Unit the management of critically ill patients is facilitated by continuous monitoring of physiological parameters and interventions to support and restore alterations of physiological functions, while laboratory and clinical data are recorded in patients medical record. All monitoring and life support devices are electronic, thus creating a large amount of data which describe patients clinical course. These bio-signals are now presented independently by each device, and analyzed only by the team of ICU physicians and nurses. The challenges for data analysis in the case of ICU are multiple, some of them being:

Analysis of vital signs variability: Normally most vital signs, such as heart rate and temperature, vary over time and this oscillatory form, which is normal, can be lost in disease states. Simple visualization of monitoring devices by doctors or nurses cannot detect the presence or absence of normal oscillations, while computer analysis can easily do so. The clinical significance of loss of normal oscillations has not been analyzed yet, although indirect evidence indicate it is highly significant.

Early detection of instability: Complications arising during the recovery period of the initial insult are not rare in critically ill patients, and significantly affect patients morbidity and mortality. Such complications, as for example septic shock from an indwelling catheter infection, are diagnosed when significant alterations in patients vital signs and laboratory tests are present. Both the normal vital signs variability and the presence of multiple physiological compensatory mechanisms, often delay the diagnosis of complications, increasing their severity. It is quite likely that subtle changes occur in the body before the compensatory mechanisms are overwhelmed, which could be detected by complex analysis of patient data.

Detection of poor patient ventilator interaction: Normally the act of breathing, although unconscious, is under strict control of the central nervous system, and forced alterations of breathing pattern, such as an unexpected occlusion of the airway, produce dramatic discomfort. In patients with respiratory failure, who require mechanical ventilation, the interaction with the ventilator is very important for patient

comfort and recovery. Yet the tools available for the clinicians to monitor patient ventilator interactions are very limited. A prototype device available in AEGLE partner PAGNI, called the PVI, has been evaluated for the analysis of patient ventilator interaction. The PVI produces a large amount of data in a format that cannot be applied at the bedside. Yet analysis of PVI data could provide information in a user-friendly format and facilitate management of patient ventilator interactions by the clinicians.

Evaluation of patients effort on mechanical ventilation: The goal of assisted mechanical ventilation is to provide some, but not all of the work required for breathing. Too little assistance would result in patients discomfort and stress, while too much assistance would result in weakening of the respiratory muscles from lack of use. At the moment evaluation of patients effort requires special monitoring devices and complicated calculations, only performed in an asynchronous manner. The data available to AEGLE from the PVI monitor provide important information on patients effort and could be used to identify how patients effort could be estimated from ventilator-derived parameters. Any improvement in patient management in intensive care has very important implications for both patients and the health care system. The ICU beds are the most expensive in the hospital, and their limited availability places in great risk any patient in need. Every day in an ICU costs over 1000 euros, and prolonged ICU stay is associated with increased morbidity, mortality and worse long term outcomes, such as cognitive dysfunction and physical impairment.

2) Analysis of data characteristics: The data generated in the ICU have the typical characteristics of big data generated by complex systems. As multiple organ-system functions are continuously monitored, streaming data of large volume are produced, including blood pressure, heart rate, temperature and others. Additionally in most patients, breathing is assisted and so mechanical ventilation provides continuous respiratory support, and also generates streaming data on respiratory system function and patient-ventilator interaction. Several therapeutic interventions are given as continuous infusions, and the rate of infusion, which is recorded by electronic pumps, provides important information on patients condition. Finally clinical and laboratory data accumulate during the hospitalization of the patient. Overall, for every single critically ill patient, a large amount of data is generated during the course of disease. Most of these data represent continuous recordings that require rapid analysis and responses. The source of data for any patient is diverse, as it includes vital signs monitoring devices, ventilators and infusion pumps, as well as medical records and laboratory results. The result is variability in the type and format of the generated data. Added complexity in data analysis results from missing data, poor input quality due to monitoring device dysfunctions or disconnections, and artifacts.

3) Limitations of existing approaches suggesting Big Data Analytics: So far in the field of Intensive Care Medicine the use of computer-assisted diagnostic and decision-supportive systems is limited. Yet the amount of data produced for every single patient is very big, exceeding any physicians capabilities, and thus underscoring the need for more sophisticated analytics. The complexity of the systems and the presence of multiple interconnections necessitate application of big data analytics approaches. Normal or abnormal variation of vital signs, and early deterioration cannot be yet detected with the means currently available in clinical practice. Thus, ICT tools such as those proposed by AEGLE, which can handle fast-changing multi-dimensional variables, can be used for the detection of abnormal variability or instability. Moreover patient ventilator interaction remains a largely unexplored area in intensive care, due to the complexity of the signals generated and the lack of appropriate means for analysis. In this case too, big data analytics approaches, as proposed by AEGLE will facilitate a better description of patient ventilator interaction, and provide physicians with valuable tools to improve patient management.

VI. Technologies Enabling AEGLE Realization

A. Big Data Frameworks

1) Hadoop Framework: The Hadoop Distributed File System (HDFS) [19] has been developed by Apache, as part of the Apache Hadoop Core project [10]. Applications that run on HDFS have large datasets, meaning that a typical file in HDFS is gigabytes to terrabytes in size. Thus, HDFS is tuned to support large files. It provides high aggregate data bandwidth and scales to hundreds of nodes in a sin- gle 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.

The supported access pattern for the files that are stored in HDFS is mutation by appending new data rather than overwriting existing data or writing data at a random offset. This assumption greatly simplifies coherency issues and places the focus of performance optimization on the append operation. In addition, data reads are sequential in most cases. Applications that run on HDFS need streaming access to their datasets. They are not general purpose applications that run on general purpose file systems. As a result, HDFS is designed for batch processing rather than interactive use by users. The emphasis is on high throughput of data accesses rather than low latency of data accesses. HDFS provides interfaces for applications to move themselves closer to where the data is located because of the fact that a computation requested by an application is much more efficient if it is executed near the data it operates on, especially when the size of the data is huge. This minimizes network congestion and increases the overall throughput of the system. The assumption is that it is often better to migrate the computation closer to where the data is located rather than moving the data to where the application is running.

2) MapReduce Programming Paradigm: The MapReduce paradigm [20] 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

[10]https://hadoop.apache.org

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. Therefore, it provides a level of abstraction that hides the messy details of parallelization, fault tolerance, data distribution and load balancing, allowing programmers to express the simple computations that they are trying to perform.

B. Medical Data Anonimization

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 consortium 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[11], 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) [21].

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 [22]. 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. It can do so on metadata such as DICOM file headers, or even raw data like DICOM images, case report forms (CSV, XML etc), HL7 messages and genetics related formats. 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. Over the last 7 years, FedEHR matured from its application in diverse fields of medical science [23], from advanced biomedical research, to translational medicine and clinical. It was thus installed, tested and used with a total of 5M EHR from 10 hospitals internationally.

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. Big Data Analytics problems are becoming increasingly common in a range of application areas such as web search and social media, finance, engineering, biology, and medical research. Ever increasing data volumes have lead to the development of a number of new parallel processing models such as MapReduce. However, in the last years, data volumes have increased at a faster pace then the available processing power, thus making it increasingly difficult to keep up with the processing requirements of modern Big Data Analytics applications. 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[12] commercializes this approach in its high-performance multiscale dataflow computing technology. Maxelers 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$\times$ 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

[11] www.gnubila.fr

[12] http://www.maxeler.com

algorithmic level, customized dataflow engines (DFEs) will be explored [24] 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 [25], [26] 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 requirements as well as lower cost than standard server systems.

VII. Standards and Privacy Aspects

A major aim of the project is to standardise databases for the purposes of wider and more effective utilisation across Member States. This ultimately requires a consolidation of laws, regulations, ethical approaches and standards. The standards applicable to Big Data will be reviewed and compared to ensure how these meet compliance to existing and established regulations in the different healthcare systems and how these may compare across healthcare systems and when dealing with medical databases. Compliance will incorporate HIPAA security access to the system, the use of Reference Information Model (ISO/HL7 21731), the adoption of CDA (ISO/HL7 27932) for the exchange of messages, the use of standardised EHR (ISO 18308) and the realization of pseudo-anonymisation techniques (ISO 25237) for the protection of sensitive personal data. The risk management based on ISO 14971 will be a horizontal action throughout the realization of AEGLE framework.

Legal issues related to privacy and data protection and how these are applied in different Member States will need to be reviewed and compared with similarities and differences clearly identified. The important issues will be the role and responsibilities of the Data Controller, the definition of sensitive personal data, the methods of anonymisation, the type and duration of security applied, the implementation and purpose of data transfer, the duration and purposes for which data is held. Ethical requirements will need to be established from the fundamental ethical principles and taking into account how ethics is currently defined and applied in each member state. The latter is novel as to date there is no clear ethical framework for dealing with Big Data within the auspices of medical databases.

VIII. Conclusions

In this paper, we presented the AEGLE approach for 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. The paper focused its analysis on the advanced Big Data use cases addressed as well as the key enabling technologies to be utilized within AEGLE project.

References

[1] Transforming health care through big data. Available online at http://www.ihealthtran.com/.

[2] 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.

[3] Top Big Data opportunities for health startups. Available online http://healthstartup.eu/2012/05/top-big-data-opportunities-for-health-startups/.

[4] Milena Sant, Claudia Allemani, Carmen Tereanu, Roberta De Angelis, Riccardo Capocaccia, Otto Visser, Rafael Marcos-Gragera, Marc Maynadié, Arianna Simonetti, Jean-Michel Lutz, Franco Berrino, and . Incidence of hematologic malignancies in europe by morphologic subtype: results of the haemacare project. 116(19):3724–3734, 2010.

[5] Chiorazzi N et al. N engl j med 2005;352:804-815.

[6] Scarfo and Ghia. Hematol oncol clin north am. 2013;27:251-65.

[7] Villamor N et al. Semin hematol. 2013;50:286-295.

[8] Sutton L et al. Semin cancer biol. 2013;23:399-409.

[9] Hallek M et al. Blood 2008; 111:5446-5456.

[10] Lafeuille MH et al. Leuk lymphoma. 2012;53:1146-1154.

[11] Stevenson FK et al. Semin hematol. 2014;51:158-167.

[12] Hallek M. Am j hematol. 2013;88:803-816.

[13] Sutton L et al. Semin cancer biol. 2015 may 8.

[14] Diabetes UK. Awareness Campaign.

[15] NHS Choices. Diabetes type 2-causes of type 2 diabetes.

[16] National Institute for Health, Care Excellence. Preventing type 2 diabetes: risk identification, and interventions for individuals at high risk. NICE public health guidance 38 [Online] 2012. Available from: http://www.guidance.nice.org.uk/ph38 [Accessed 25th October 2013].

[17] Griffo E Rivellese A. Giacco R, Vetrani C. Role of diet and diet interventions in diabetic patients: Physiological and metabolic changes and reduction in morbidity and mortality. Current Nutritional Reports. 2013; 2: 174-180.

[18] Holman RR et al. Adler AI, Stratton IM. Association of systolic blood pressure with macrovascular and microvascular complications of type 2 diabetes (ukpds 36): prospective observational study. BMJ. 2000; 321:412419.

[19] 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.

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

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

[22] et al. Berlanga R., Manset D. 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.

[23] A. Gaignard N. Boujelben S. Gaspard et al. S. Cipire, G. Ereteo. Global initiative for sentinel e-health network on grid (ginseng), medical data integration and semantic developments for epidemiology cluster, cloud and grid computing (ccgrid), 2014 14th ieee/acm international symposium. pages 755-763. 02/2014.

[24] 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.

[25] S. Xydis, A. Bartzas, I. Anagnostopoulos, D. Soudris, and K. 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.

[26] Sotirios Xydis, Ioannis S. Stamelakos, Alexandros Bartzas, and Dimitrios Soudris. Runtime tuning of dynamic memory management for mitigating footprint-fragmentation variations. In *ARCS 2011 - 24th International Conference on Architecture of Computing Systems 2011, Workshop Proceedings, February 22-23, 2011, Como, Italy.*, 2011.