
Proceedings Paper:
de Vila, MH, Attar, R, Pereanez, M et al. (2019) MULTI-X, a State-of-the-Art Cloud-Based Ecosystem for Biomedical Research. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018 IEEE (BIBM), 03-06 Dec 2018, Madrid, Spain. IEEE, pp. 1726-1733. ISBN: 978-1-5386-5488-0. ISSN: 2156-1125. EISSN: 2156-1133.

https://doi.org/10.1109/BIBM.2018.8621317

MULTI-X, a state-of-the-art cloud-based ecosystem for biomedical research

Milton Hoz de Vila[1], Rahman Attar[1], Marco Pereanez[1] and Alejandro F. Frangi[1,2]

[1]*Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK*
[2]*Biomedical Imaging Department, Leeds Institute for Cardiovascular and Metabolic Medicine (LICAMM), School of Medicine, University of Leeds, UK.*

***Abstract*— With the exponential growth of clinical data, and the fast development of AI technologies, researchers are facing unprecedented challenges in managing data storage, scalable processing, and analysis capabilities for heterogeneous multi-sourced datasets. Beyond the complexity of executing data-intensive workflows over large-scale distributed data, the reproducibility of computed results is of paramount importance to validate scientific discoveries. In this paper, we present MULTI-X, a cross-domain research-oriented platform, designed for collaborative and reproducible science. This cloud-based framework simplifies the logistical challenges of implementing data analytics and AI solutions by providing pre-configured environments with ad-hoc scalable computing resources and secure distributed storage, to efficiently build, test, share and reproduce scientific pipelines. An exemplary use-case in the area of cardiac image analysis will be presented together with the practical application of the platform for the analysis of ~20.000 subjects of the UK-Biobank database.**

***Index Terms*—biomedical informatics, precision medicine, cloud computing, population analysis.**

I. Introduction

By discovering the connections of patient's unique genetic, biological, environmental, and lifestyle characteristics, precision medicine intends to determine how these individual aspects affect the evolution of the diseases and aims to develop tailored prevention and treatment strategies. Medical imaging has become a primary tool for such analytic approach towards personalised medicine. Non-invasive imaging data, together with a vast range of highly heterogeneous non-imaging patient health information ground the concept of *medical big data*. Machine learning (ML), as an emerging key discipline in data science, offers the unprecedented capacity to exploit and generate knowledge from such large, disparate, and multidimensional datasets. However, in biomedical research, tackling the computational complexity related to performing medical big data analysis requires a deep understanding of its particular technical requirements and limitations [1]. Even with the leverage of sophisticated cloud computing services, implementing big data analysis tools and high-throughput solutions is still highly complex, expensive and impractical in many cases [2].

II. Current Challenges

The implementation of precision medicine solutions, supported by Artificial Intelligence (AI) technologies and applied over medical big data is at a nascent stage [3, 4]. Several challenges must be addressed, not only to satisfy the requirements of such extreme-scale data analytics [5] but also, to leverage the full potential and capabilities of a growing and highly active community of scientist coming from different interdisciplinary fields. Besides the essential computational prerequisites, such as performance, capacity, and reliability, the most crucial challenges are related to the integration of many key practical, technical, and non-technical aspects. Here, we outline those specific challenges that a state-of-the-art research platform for precision medicine must address.

Scalability. To reach the extreme-scale level in data processing, analytic solutions need to be able to coordinate a large number of concurrent service requests and automatically provide resources to satisfy the computational demands.

Reproducibility. The ability to reproduce *in-silico* experiments is recognised to be a major need to validate and also verify the discoveries [6], and furthermore, to allow reusability. To enable replication, all the involved aspects need to be considered and reproduced: data, applications, computational resources, operating conditions and execution environment.

Flexibility. Developers and data scientist produce invaluable material to enable scientific discovery. Methods, algorithms, analytics applications, models, visualisation tools, curated datasets, and other assets are being continuously generated and are made available to the broader community, many times with open access license terms. To leverage such heterogeneous resources, precision medicine frameworks need to support multiple programming environment and be capable of integrating new components as they become available regardless of their programming languages, operating systems, or hardware specifications requirements.

Interoperability. As outlined by the FAIR principles [7], research requires data to be Findable (citable), Accessible (trackable), Interoperable (intelligible) and Re-usable (reproducible). Several regional and global initiatives have emerged to support this approach. Open data repositories, open-

source applications and open access literature are freely available through those infrastructures. A modern research platform must be able, not only to take advantage of these resources but also be capable of exposing new resources and making them reachable and interoperable. Standards for APIs and workflow protocols, archetypes and ontology development are essential components to build such interoperability services.

Fast Prototyping. A user-friendly framework that simplifies operations and consistently integrates different components is crucial to develop, test, and efficiently deliver research solutions. Emerging container-based virtualisation frameworks can be used to reduce complexity and accelerate distribution, testing and deployment mechanisms.

Workflow Management. The automation of the execution of computational tasks is a key factor to improve scientific productivity [8]. Workflows provide the interface between scientist and complex computing infrastructures. Extreme-scale workflow managers need to deliver sophisticated orchestration capabilities to operate efficiently in a complex and highly dispersed computing model.

Data Management. Biomedical scientists face major difficulties in finding, using and publishing datasets. Modern data management systems must provide flexible and feature-rich solutions to address present-day challenges: (i) *data quality assessment*, offer mechanisms to verify completeness, accuracy and consistency of the data processed; (ii) *data heterogeneity*, enable a highly adaptable and scalable storage architecture for the management of multi-scale and multidisciplinary datasets; (iii) *big data warehousing*, provide efficient petabyte-scale storage and querying capabilities for structured and non-structured data across heterogeneous systems; (iv) *data federation*, deploy distributed file systems across diversified storage architectures facilitating local/global multi-location data access using unique identifiers and permissions; (v) *data homogenisation*, integrate semantic components, enable metadata annotation, semantic search, and semantic mediation to tackle any mismatches in the data or their format; (vi) *data privacy*, enable user interfaces to review and manage data usage agreements, granted permissions, and to inspect compliance with institutional policies and other legal considerations; (vii) *data security,* employ International Standards Organization (ISO) certified protocols and system providers to establish a strong IT security strategy. (viii) *data access,* provide a unified identity management system to enable single sign-on services and granular permissions control across distributed systems and services; (ix) *data publication*; programmatically assign Data Object Identifiers (DOI), and facilitate the exposure and linkage of each data asset; (x) *data interoperability*, facilitate the injection and integration of external data sources, building on top of the existing standards in health and research domains [9], such as OpenEHR archetypes and CEN EN 13606, the European Patient Summary, HL7 C-CDA templates, HL7 FHIR Resource Profiles, CIMI archetypes, BRIDG Domain model, and the CDISC standards.

Data visualisation. Visual analytics is an overarching factor that facilitates knowledge discovery. Similar to any other research areas, visualisation techniques need to be adapted to the characteristics of the big data paradigm [10]. Explore, aggregate, manipulate, analyse, and visualise highly dispersed healthcare datasets and correlating them to derive insights remains a challenge.

Licensing and IP Management. During the last years, we have witnessed the exponential growth of big data generation and availability and also, great advances in ML and AI technologies. Most of these solutions are cloud-native or are easily portable to a cloud or virtual environment. The sharing, distribution, and licencing of such disparate assets introduce an unprecedented level of complexity. Practical mechanisms must be implemented to facilitate the discovery and improve the leverage of these resources. Research frameworks have to be able to manage multiple data sources, software and service providers, enabling flexible handshaking protocols to access open and proprietary solutions, managing diverse licensing models and usage agreements.

Cost Assessment. Budget estimation, allocation and monitoring are essential components of conducting scientific activities. Research platforms need to provide scientist advanced cost evaluation mechanisms which will have a direct impact on the scope and viability of their projects.

Security and Compliance. IT security mechanism must be guaranteed using international standards. Furthermore, legal and regulatory aspects must be assessed and accomplished. Covering these mandatory areas require the involvement of experts in each field. Satisfying these requirements is challenging for small, and even for large organisations.

Collaboration. Biomedical research depends upon the integration of different fields, methodologies and technologies. A flexible framework to collaborate integrating multidisciplinary teams and their multi-field solutions provides a comprehensive ecosystem to underpin scientific research.

III. Related Work

The working environment for researchers and scientific developers is quickly evolving from isolated institutional High-Performance Computing (HPC) clusters to a global ecosystem of interconnected open cloud platforms. Exploiting the potential of scientific data and applications through internet-accessible services is now a common approach for developing research solutions. Many initiatives have been introduced to make this possible.

Comprehensible workflow execution systems have emerged providing versatile and efficient solutions. Among the most representative are Taverna, Pegasus, Kepler and KNIME, offering traditional desktop-based applications to create, edit, execute and share scientific workflows. Likewise, other exceptional tools, such as Swift, Nipype [11], Hyperflow, MOTEUR, Triana, VisTrails, Apache Airflow and GIMIAS, were built to support scientific pipelines in a variety of research areas. In all cases, the integration of these solutions with data components and computational resources varies according to each specific implementation. This variability of execution environments and data components is a major issue for reproducibility [6], scalability and collaboration.

Other comprehensive development platforms, such as

Galaxy [12], LONI [13], SADI and InterMine have been conceived with a significant emphasis on data integration. They are rather highly specialised frameworks. LONI, for example, is an exceptional resource for neuroimaging applications. Galaxy provides built-in libraries with a major focus on computational biology but is rather unsuitable for storage-demanding medical imaging pipelines. They are massively used in their specific domains, although they often lack flexible models to integrate new components, or solutions for scalable computational resources.

With a similar approach, Science Gateways [14] such as CBRAIN, NSG, CSGF, WS-PGRADE/gUSE, and Globus Galaxies, have been deployed to integrate efforts and share existing resources. These systems allow programmatic access to external data sources and include one or several workflow engines. Apache Airavata [15], is a software framework used by many scientific gateways to manage and execute workflows, it provides an abstraction layer to deploy tasks across distributed computing resources, via a ReST API. The complexity associated to execute and integrate new components on these platforms depends on their integration model and the instantiation type used in each case. For instance, Globus Galaxies, which is built upon Globus Genomics and Galaxy, delivers a domain-independent framework with a flexible cloud-based pay-as-you-go service. The integration model used to deploy and execute new applications in this framework requires the development of specific wrappers and contextualisation recipes.

Other notable cloud-based integrative approaches are the iPlant Collaborative [16] framework, which offers a comprehensive cyberinfrastructure for life science research; and the American Heart Association Precision Medicine Platform[1], which provides openly accessible datasets and state-of-the-art tools in the cardiac field. Emerging examples in the private sectors are Philips HealthSuite Insights[2], which aims to deliver the first AI ecosystem for the healthcare industry, it offers a comprehensive environment to build, share and deploy decision support solutions; and InSilicoTrials[3], a collaborative framework for computational modeling and simulation primarily targeted to pharmaceutical and medical devices development.

Although many of these frameworks are robust and integrative solutions, and can be deployed on scalable cloud environments, the way they handle data, computing resources, as well as their collaborative approaches, differs strongly from the comprehensive, modular and integrative strategy devised in this work. Our position is that most of the existing platforms lack interoperability between extreme-scale applications, computational resources, and heterogeneous datasets. Frameworks often lack usability, requiring often a cumbersome training phase before starting to work on them. Moreover, a global oversight of other essential functional aspects, such as those described in the previous section, are mostly unattended or insufficiently covered.

[1] https://precision.heart.org
[2] www.healthsuiteinsights.philips.com
[3] https://insilicotrials.com

IV. Implementation

A. Overview

MULTI-X (https://multi-x.org) is a cross-domain research-oriented platform, designed for collaborative and reproducible science. It provides a general-purpose, scalable, and integrative computational framework. It was designed to foster development, integration, and testing of scientific tools, facilitating their connection with large-scale federated data repositories and providing analysis tools to get immediate insights about the data processed. This comprehensive and modular framework relies on the interoperability of 6 differentiated areas as depicted in Fig. 1. :

1) *DATA,* a flexible repository of annotated data collections for structured and non-structured data.
2) *ANALISE,* a catalogue of multi-domain scientific tools fed by a flexible development environment based on containers and virtual machines.
3) *COMPUTE,* cloud-agnostic middleware providing efficient and scalable on-demand computational resources.
4) *WORKFLOW,* a set of services for the composition, orchestration, schedule, and automated execution of data analytics processes.
5) *EXPLORE,* analysis dashboards and visualisations tools providing coordinated representation and exploration of highly multivariate data.
6) *COLLABORATE,* web-based user-friendly interfaces with fully managed sharing capabilities and functional components to streamline interaction within the various internal systems, and to enable interoperability with external infrastructures.

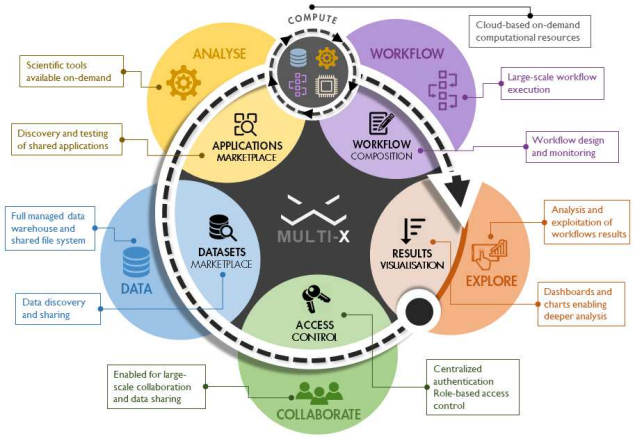

Fig. 1. Diagram of the MULTI-X components and the platform key capabilities. The user's process flow to produce or reproduce scientific results in MULTI-X typically involves the following steps: a) the access to a dataset, query, and selection of a subset; b) the deployment of an analysis tool; c) the use of a workflow engine to execute the tool developed over a large dataset; d) the visualisation of the results obtained in the previous step; e) the sharing of the results, tools or workflows generated.

B. Architecture

MULTI-X is a Research as a Service (RaaS) platform. It is based on the Platform as a Service (PaaS) model, and a

Distributed Operating System (OS) architecture. PaaS is a category of cloud computing services which allow users to develop, run, and manage applications without the complexity of building and maintaining an IT infrastructure. PaaS services are hosted in the cloud and accessed by users using a web browser. On the other hand, Distributed OS are sophisticated infrastructure abstractions were multiple services, applications, and computational resources are orchestrated from a single interface, as a single distributed OS (Fig. 2).

The platform takes advantage of 3 core technologies, (i) Infrastructure as Code (IaC), which refers to the provisioning of computing resources based on definition files; (ii) Cloud Computing, which enables ubiquitous access to scalable computational resources; (iii) Containers Orchestration, a flexible approach to manage hardware-agnostic virtualisation by accessing a shared operating system kernel.

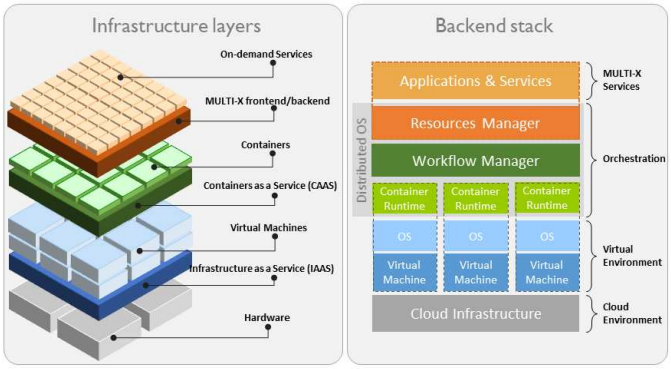

Fig. 2. The MULTI-X modular approach is underpinned by a flexible multi-layer virtualised infrastructure (left), and a cloud-native backend distributed system for service orchestration (right).

C. Specifications

Below is a brief description of how MULTI-X addresses the challenges outlined in Section II.

Scalability. The platform provides a highly-configurable execution engine for efficiently loading cloud computing resources. Different processing mechanisms can be defined to cope with varying workloads. Locally, in series or parallel, or distributed, using load-balanced dynamically generated computing clusters.

Reproducibility. Full execution environments can be saved, shared and reproduced. Data provenance, applications parameters, workflow configuration and hardware instantiation methods, are stored for further development, validation or reuse.

Flexibility. The most widely used programming languages and frameworks such as C/C++, C#, Java, Perl, R, Python, or Ruby are supported for the development of applications and also for workflow composition. Other open-source or commercial tools such as Matlab, Abaqus or Ansys can be easily incorporated using appropriate licenses. ML (including Deep Learning) libraries and frameworks with GPU support are also available on pre-installed environments.

Interoperability. A flexible and extensible API framework is used to interoperate with external systems and allow the integration of new functionalities. Although still in a development stage, an RDF triple store based on HBase and existing domain ontologies will be integrated to enable semantic association discovery.

Fast Prototyping. Development environments can be shared, instantiated and saved on different stages of the development process. With a similar approach, Docker and Singularity containers can be created or imported, and used for workflow composition and execution.

Workflow Management. The platform supports multiple state-of-the-art workflow engines such as Nipype, NextFlow and Taverna, providing enhanced mechanisms to orchestrate virtual, cloud and containerised environments. Comprehensive templates are available in different programming languages to facilitate the creation of cloud-based computing clusters and to execute computational pipelines from a user-friendly web interface or the command line.

Data Management. For structured data, a fully managed columnar database cluster enabled for big data analysis is used for petabyte-scale data warehousing service. For non-structured data, a cloud-based distributed Object File System is made available through the integrated Rule-Oriented Data System (iRODS) which provides secure multi-protocol access using API, WebDAV or REST invocations. This Shared File System provides a reliable, robust and secure virtual file store capable of interconnecting all the pieces of information required for the execution of workflows, the analysis and visualisation of data, the extraction and management of metadata, and the implementation of policies for data control in the cloud environment. The underlying Identity and Access Management (AWS-IAM) system enable unified access and tight integration with other modular extensive components such as the MULTI-X Data Collection, Data Injection, and Data Harmonisation services.

Data visualisation. Predefined dashboards can be configured to showcase the results of the experiments enabling immediate analysis of the data processed. Graphs, charts, images, tables, 2D and 3D viewers can be accessed from the web interface using secure sharing protocols based on JSON Web Tokens (JWT). To enhance dissemination, Jupyter notebooks can be linked to each asset created in MULTI-X.

License and IP Management. A genuine marketplace is provided for sharing and distribution of data, applications, and full execution pipelines. The entire process of licensing and handshaking is supported and monitored. The system offers commonly used licensing models to choose from and editable templates for electronic agreement of Terms and Conditions.

Cost Assessment. Users and organisations can bring their own cloud accounts; they can monitor expenditures, calculate the cost of large workflow executions, and explore options to reduce charges. Currently, Amazon Web Services (AWS) is fully supported, although other public cloud providers and OpenStack-based frameworks can be incorporated.

Security and Compliance. Industry-level security protocols are integrated into all the different layers to guarantee restricted access only to authorised users, and to safeguard the transfer of information between all components using a centralised Identity and Access Management service and ISO certified cloud infrastructures.

Collaboration. Each asset deployed in the platform can be shared, and its usage monitored. The whole ecosystem has been designed for secure and controlled sharing, connecting people and organisations, enabling reproducibility, and making efforts available to the research community.

V. Use Case: Automatic biventricular segmentation and quantification of the UK Biobank population cohort

A. Scope and limitations

The effectiveness in addressing each of the challenges identified in Section II need to be considered in evaluating the level of usability and the overall efficacy of a research-oriented infrastructure such as MULTI-X. Although a full assessment would require a more detailed and extensive analysis, this paper focuses on two key relevant aspects.

1) Demonstrate the applicability and practicality of the platform to deploy and reproduce high-throughput computational and data-intensive analysis using cost-effective scalable resources.
2) Expose the knowledge discovery capabilities of the platform by combining structured and non-structured data, integrating imaging and "omics" data with demographic, clinical, and patient-reported data, such as lifestyle factors and diet.

B. Overview

As an exemplary use case, a novel high-throughput pipeline in the area of cardiac magnetic resonance (CMR) image analysis was deployed in MULTI-X. Subsequently, employing the method incorporated, the platform was used to process a large scale cohort study (a population of 20,000 subjects imaged at 50 cardiac phases each, i.e. on 1 million CMR volumes), derived from the UK Biobank [17]. Finally, clinical and demographic data were integrated with the results and used to enrich the representation and the interpretation of the experiment.

C. Analysis tools and workflows

The analysis of cardiac structural and functional indexes in large-scale population-based studies can reveal patterns and linkages across population sub-groups, and provide data- driven insights into risk factors before cardiovascular diseases (CVDs) occur. In a very recent publication [18], Attar and colleagues presented a fully automatic workflow capable of performing end-to-end 3D cardiac image analysis. This pipeline encompasses several individual processing steps (Fig. 3), it was developed using different ML based methods in diverse programming languages and seamlessly integrated into a Nipype pipeline.

The method is capable of segmenting the cardiac ventricles and generating global and regional clinical reference ranges that are statistically comparable to those obtained by human observers. Its clinical impact derives from the 3D analysis of left ventricle (LV) and right ventricle (RV) of the heart, as well as the extraction of key cardiac functional indexes from large CMR datasets. The following indexes are calculated from the automated segmentation and then compared to measurements from manual segmentation: LV end-diastolic volume (LVEDV), end-systolic volume (LVESV), LV stroke volume (LVSV), LV ejection-fraction (LVEF), LV myocardial mass (LVM), RV end-diastolic volume (RVEDV) and end-systolic volume (RVESV), RV stroke volume (RVSV) and RV ejection-fraction (RVEF).

A cohort of 800 healthy subjects selected from the UK Biobank for which manual delineations and reference ranges exist [19], was used by the authors to validate their results. The human-like reliability of the method was demonstrated by comparing the accuracy of the obtained segmentations with the error ranges observed between different human raters. Furthermore, significant statistic correlation was reported between the manually retrieved reference indexes and those automatically computed with the pipeline. From these results, it was demonstrated that the method can be used to fully automate the segmentation and quantification of large scale datasets where manual inputs are infeasible.

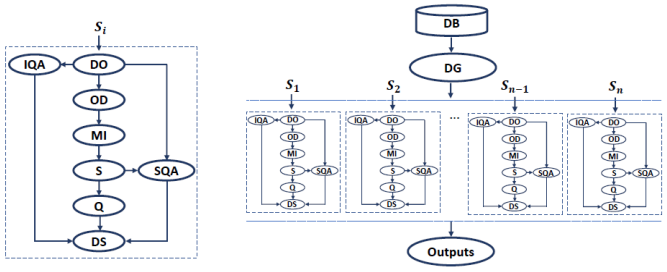

Fig. 3 R.Attar et al. fully automatic image parsing workflow for the analysis of cardiac ventricles in parallel. Left: Modules integrating the pipeline: DO: Data Organisation, IQA: Image Quality Assessment, OD: Organ Detection, MI: Model Initialisation, S: Segmentation, SQA: Segmentation Quality Assessment, Q: Quantification, DS: Data Sink. Right: The quantitative functional analysis of a large database in parallel mode. DB: Database, DG: Data Grabber, n: number of subjects, and S_i: i^{th} subject of the dataset.

In this paper, we use the same pipeline introduced by [18] and up-scale from 800 to 20,000 subjects including their clinical metadata to demonstrate the scalability and knowledge discovery of our platform.

D. Data

The UK Biobank is a large prospective epidemiological cohort study. It includes extensive phenotypic and genotypic data together with detailed health and lifestyle information on 500,000 participants, aged 40-69 years. Ongoing longitudinal follow-up studies will acquire and store additional multi-organ, multi-modality imaging data from 100,000 of the original volunteers by 2022. By the date of this study, CMR scans of 20,000 volunteers were already collected and made available for further research.

E. Implementation in MULTI-X

To accomplish the objectives defined in Section V.A, the method developed by Attar et al. [18] was deployed in MULTI-X, and 20,000 UK Biobank follow-up studies were analysed. The segmentation volumes and biventricular reference ranges for each of these subjects were obtained and stored. Furthermore, clinical and demographic information of these subjects, queried from the UK Biobank database, were used to generate data visualisation dashboards enabling a more in-depth analysis by combining the results with clinical, imaging and demographic data.

Below is a brief description of the steps that were followed to implement this use case. It also represents an overview of the essential components of the platform.

1) *Integration of the analysis tools (ANALISE component)*
From the "*ANALISE*" area in MULTI-X, a new development machine was created and launched. Using an interactive form, the operating system, essential libraries, and computational resources, particularly CPU, RAM and disk I/O capabilities, were defined to create a cloud-based instance according to the required specifications. Once initiated, the system was accessed through standard protocols, using the command line or the graphic interface. All the execution modules involved in the 3D cardiac image analysis (*Fig. 3*) together with the Nipype code were uploaded and tested in the new environment. During the implementation process, the machine was saved and stopped, and subsequently re-launched until the imported applications were fully operational.

2) *Data upload (DATA component)*
Using secure protocols and encryption mechanisms, the UK Biobank anonymised data were imported into the MULTI-X object storage backend. Thereafter, each time an instance of the development machine was launched, the CMR scans were made available as input for the executions. Subsequently, the structured data containing the Electronic Health Records (EHR) of the same follow-up study were uploaded to the core database. In the "*DATA*" section of MULTI-X, extensive information about the new dataset, including usage logs, permissions, user agreements, provenance, modalities and data types were generated and stored for monitoring and further modification.

3) *Pipeline integration (WORKFLOW component)*
A standard procedure in MULTI-X is to generate a bespoke pipeline of processing steps by editing an existing template. However, in this case, the original pipeline was developed in Nipype, a workflow framework that is out-of-the-box supported in the platform. Thus, the Nipype code was uploaded to the development machine and used straightforward with minor modifications. The execution of the 20,000 subjects was configured and launched from the "*WORKFLOW*" area of MULTI-X. The pipeline file was displayed on the web interface allowing quick modifications and immediate testing. An interactive wizard was used to define the parallelisation variables, the number and type of machines, and the location of the input and output files.

4) *Full Cohort execution (COMPUTE component)*
Preliminary tests were performed to estimate the scalability, cost and performance of the parallelisation of the full pipeline (*Fig. 5*). The execution was finally configured to launch a cloud-based cluster with 120 mid-spec machines (AWS c4.8xlarge), each one with 36 virtual CPU (vCPU) and 60 GiB RAM. The full analysis of the 20,000 subjects was completed in 2 hrs 39 min. For each subject, an average of 1400 files (650MB) were generated, including the output of the segmentation, images and VTK files for each of the 50 timepoints of the cardiac cycle, CSV files containing the functional indexes calculated by the quantification module, and other control files. In total, the pipeline generated 27M files, 10TB of unstructured data. The execution was monitored in real time from the "*COMPUTE*" area of MULTI-X, and a comprehensive set of logs, configuration files, and usage costs were stored in the history area for further evaluation.

5) *Results analysis (EXPLORE component)*
An automated script was used to parse the massive unstructured data created in the previous step. All new subjects and their cardiac indexes were extracted and uploaded to the core database to enable data visualisation. Furthermore, the UK Biobank clinical and demographic data of these anonymised subjects were queried and made available from the analysis dashboards. Hence, a simple exploratory approach was enabled to allow the identification of meaningful relationships between the cardiac function indexes derived from the automatic segmentation and the patients' clinical data. Four dashboards were created, based on the users' request, in MULTI-X: a) General Population (*Fig. 4*), for global analysis, combining clinical and demographic data and advanced filtering; b) Individual Exploration (*Fig. 6*), for subject selection and comparison; c) Parameter Optimisation (*Fig. 7*, left), to visually examine the output of the analysis by changing variables or input parameters; d) Method Evaluation (*Fig. 7*, right), to compare the results of different methods. Finally, using standard Python libraries, a Jupyter notebook was deployed to validate the analysis by representing the correlation between the manual and the automated method used to calculate the cardiac function indexes (*Fig. 8*).

Fig. 4 General Population analysis dashboard deployed in MULTI-X to interactively explore the results of the *Biventricular segmentation and quantification* pipeline. It combines the cardiac function indexes obtained from the automatic segmentation of 20,000 subjects together with the clinical, demographic and lifestyle data queried from the UK Biobank study.

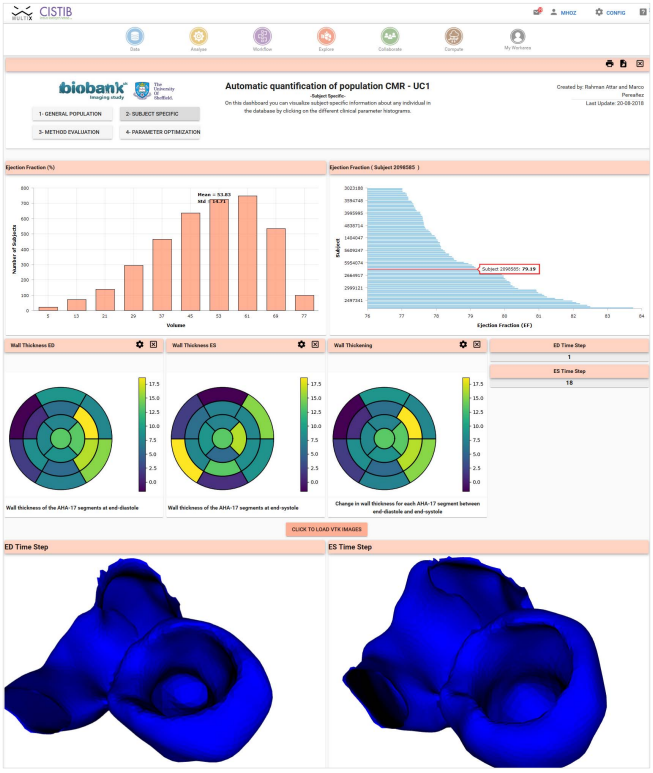

Fig. 6 Individual Exploration dashboard; by clicking on the upper-left histogram blocks, different subjects can be chosen from the upper-right selector to load the patient's specific cardiac function indexes, the AHA 17 myocardial segments, and the corresponding end-systolic and end-diastolic VTK file of 3D surface meshes generated by the pipeline.

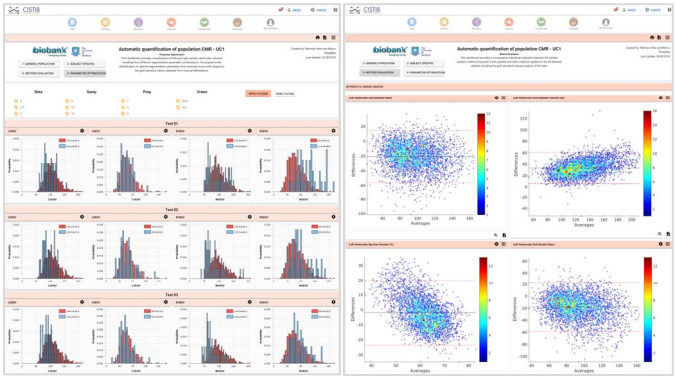

Fig. 7 Parameter Optimisation dashboard (left), an interface to visually evaluate the best values for a given set of parameters; the optimal combination is the best-matching superposition between the ground truth (red) and the automatically calculated (blue) probability histograms for cardiac indexes. *Method Evaluation* dashboard (right), Bland-Altman plots used for a comparative analysis between two methods.

F. Cost and performance

The final workflow was configured to optimise time and cost. The number of concurrent instances and the type of computing resources were defined using a setup wizard from the MULTI-X web interface. An intermediate number of parallel machines with high specifications were used to execute the pipeline (Fig. 5). The final expenditures were within the range of the initial estimates (USD 577.57). During the executions, the monitoring dashboards reported an average of 80% of CPU utilisation and memory peaks at 75% on the execution nodes indicating an efficient use of resources.

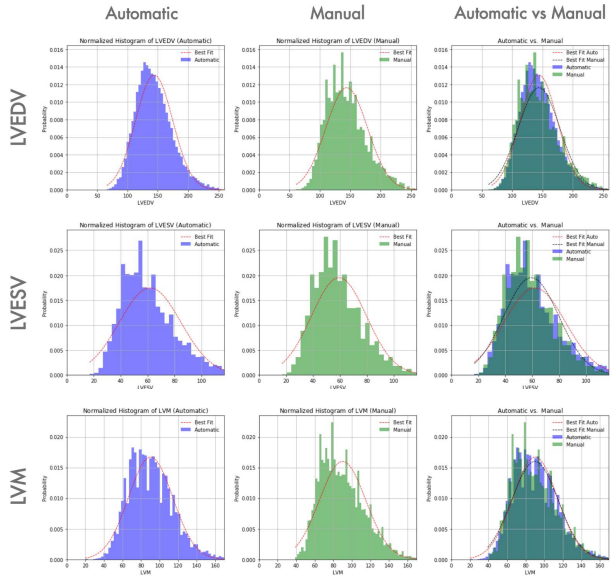

Fig. 8 Partial view of the MULTI-X Jupyter notebook deployed to evaluate the results of the workflow execution. The normalised histograms representing the cardiac function indexes automatically calculated from 20,000 subjects are significantly correlated to those manually obtained.

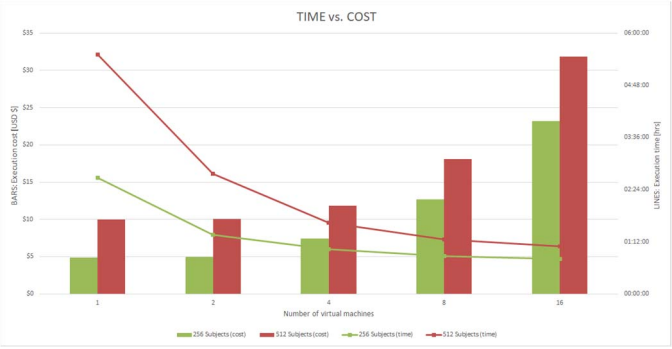

Fig. 5 Performance speedup for running a workflow in MULTI-X, with 256 and 512 subjects, on a different number of virtual machines executed in parallel. Each machine has 32 vCPU. The figure reveals an inverse correlation between the workflow execution time and the cost of running the workflow. These estimations allow selecting the most cost/time efficient option.

VI. Discussion and Conclusion

As technologies such as machine learning, big data analysis, and knowledge discovery evolve, it is essential to provide easy-to-use, time-saving solutions that make them accessible to researchers and developers. We introduce MULTI-X, a comprehensive cloud-based ecosystem tailored for biomedical research, an integrative approach to streamline the development, execution, and sharing of scientific applications and workflows.

In this article, we described the implementation of a complex pipeline which includes modeling, imaging, and machine learning techniques. To demonstrate scalability, a large cohort was analysed using high-throughput analysis workflows. Finally, combining state-of-the-art visualisation techniques and big-data querying capabilities, the data produced were associated with clinical and demographic information and exposed for further analysis into comprehensive visualisation dashboards and Jupyter notebooks.

With a focus on open science and reproducibility, the platform was designed to address emerging challenges related to the

design and deployment of multiscale/multi-domain analysis pipelines. With a modular and extensible approach, the platform is able to meet the needs of a wide range of research communities and disciplines, e.g. to support the growth of radiomics, and the promise of precision medicine.

VII. Further Work

Current and emerging technology frameworks, scientific applications, programming languages, workflow packages, and visualisation tools are systematically reviewed, tested, and integrated into the MULTI-X platform. Further use cases in the areas of genomics, data mining and semantic harmonisation will be soon deployed in the platform to demonstrate its capabilities to analyse and estimate genotypic and phenotypic correlation.

Furthermore, MULTI-X is being used as the underlying integration framework for the European funded Back-UP project which aims to produce patient-specific multifactorial prognostic models to support an effective and efficient management of non-specific Neck and Low Back Pain (NLBP). The platform will facilitate collaboration across diverse research communities and disciplines, integrating data and models derived from biological, psychological, social and economic fields.

The massive development of AI and big data technologies applied to healthcare and life sciences let us envisage a landscape where experiments of unprecedented size and complexity will be conducted in scalable computing environments. Similarly as what it happens with open data and open publication initiatives, the availability of open tools and workflows in a shared environment will boost collaboration and promote the reusability of research assets.

VIII. Acknowledgment

This work was supported in part by the European Research Council (Back-UP, ID 777090). Our thanks to Amazon Company for funding this research through AWS Research Grant (2018-2019).

IX. References

[1] M. T. Krieger, O. Torreno, O. Trelles, and D. Kranzlmüller, "Building an open source cloud environment with auto-scaling resources for executing bioinformatics and biomedical workflows," *Future Generation Computer Systems,* vol. 67, pp. 329-340, 2017.

[2] P. Ping, H. Hermjakob, J. S. Polson, P. V. Benos, and W. Wang, "Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine," *Circulation Research,* vol. 122, pp. 1290-1301, 2018.

[3] J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, "Big data analytics to improve cardiovascular care: promise and challenges," *Nature Reviews Cardiology,* vol. 13, p. 350, 2016.

[4] C. Trinh-Shevrin, R. Sacks, J. Ahn, and S. Y. Stella, "Opportunities and challenges in precision medicine: Improving cancer prevention and treatment for Asian Americans," *Journal of racial and ethnic health disparities,* vol. 5, pp. 1-6, 2018.

[5] R. F. da Silva, R. Filgueira, I. Pietri, M. Jiang, R. Sakellariou, and E. Deelman, "A characterization of workflow management systems for extreme-scale applications," *Future Generation Computer Systems,* vol. 75, pp. 228-238, 2017.

[6] S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y. L. Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, and C. Blanchet, "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities," *Future Generation Computer Systems,* vol. 75, pp. 284-298, 2017/10/01/ 2017.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, and P. E. Bourne, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific data,* vol. 3, 2016.

[8] M. Atkinson, S. Gesing, J. Montagnat, and I. Taylor, "Scientific workflows: Past, present and future," ed: Elsevier, 2017.

[9] G. B. Laleci and A. Dogac, "A semantically enriched clinical guideline model enabling deployment in heterogeneous healthcare environments," *IEEE Transactions on Information Technology in Biomedicine,* vol. 13, pp. 263-273, 2009.

[10] J. W. Dunn, A. Burgun, M.-O. Krebs, and B. Rance, "Exploring and visualizing multidimensional data in translational research platforms," *Briefings in Bioinformatics,* vol. 18, pp. 1044-1056, 2017.

[11] K. Gorgolewski, C. Burns, C. Madison, D. Clark, Y. Halchenko, M. Waskom, and S. Ghosh, "Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python," *Frontiers in Neuroinformatics,* vol. 5, 2011-August-22 2011.

[12] E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, and B. A. Grüning, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic acids research,* vol. 46, pp. W537-W544, 2018.

[13] D. E. Rex, J. Q. Ma, and A. W. Toga, "The LONI pipeline processing environment," *Neuroimage,* vol. 19, pp. 1033-1048, 2003.

[14] T. Glatard, M.-E. Rousseau, S. Camarasu-Pop, R. Adalat, N. Beck, S. Das, R. F. da Silva, N. Khalili-Mahani, V. Korkhov, and P.-O. Quirion, "Software architectures to integrate workflow engines in science gateways," *Future Generation Computer Systems,* vol. 75, pp. 239-255, 2017.

[15] M. E. Pierce, S. Marru, L. Gunathilake, D. K. Wijeratne, R. Singh, C. Wimalasena, S. Ratnayaka, and S. Pamidighantam, "Apache Airavata: design and directions of a science gateway framework," *Concurrency and Computation: Practice and Experience,* vol. 27, pp. 4282-4291, 2015.

[16] N. Merchant, E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Micklos, and P. Antin, "The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences," *PLoS biology,* vol. 14, p. e1002342, 2016.

[17] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, and F. E. Rademakers, "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance,* vol. 15, p. 46, 2013.

[18] R. Attar, M. Pereañez, A. Gooya, X. Alba, L. Zhang, S. Piechnik, S. Neubauer, S. Petersen, and A. Frangi, "High Throughput Computation of Reference Ranges of Biventricular Cardiac Function on the UK Biobank Population Cohort," presented at the Statistical Atlases and Computational Modelling of the Heart Workshop, Granada, 2018.

[19] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, J. M. Francis, M. Y. Khanji, E. Lukaschuk, and A. M. Lee, "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort," *Journal of Cardiovascular Magnetic Resonance,* vol. 19, p. 18, 2017.