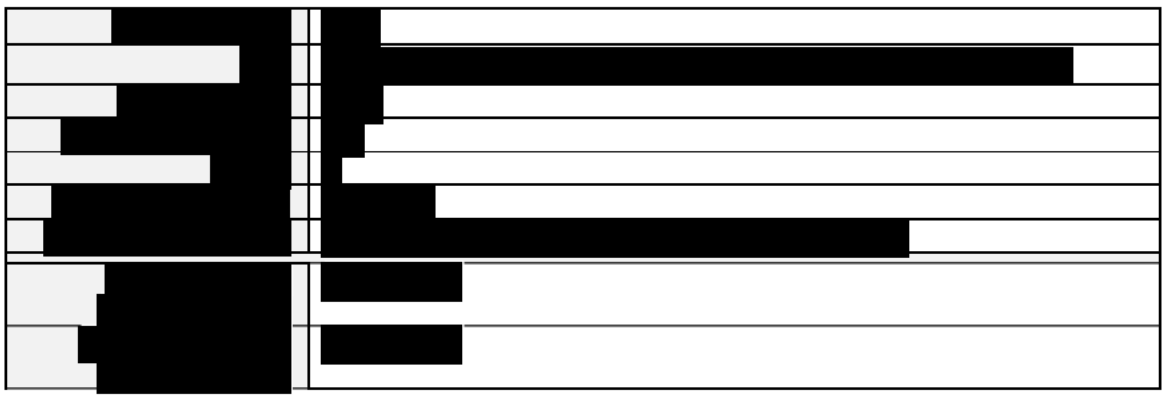
PRIMAGE

Medical imaging
Artificial intelligence
Childhood cancer research

D2.2 – Hybrid Open Datacloud and processing middleware architecture

Project Full Title: *PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers Business Place*
Project acronym: *PRIMAGE*
Project type: Horizon 2020 | RIA (Topic SC1-DTH-07-2018)
Grant agreement no: 826494

[redacted]



1. Introduction

One of the main objectives of PRIMAGE project is to develop a ***hybrid open datacloud and processing middleware*** which comprises the use of open public and private clouds, and specific processing resources, in particular High Performance Computing (HPC) and cloud, to support predictive tools for assisting diagnosis, prognosis, therapies choice and treatment follow up, based on the use of novel imaging biomarkers, in-silico tumour growth simulation, advanced visualisation of predictions with weighted confidence scores and machine-learning based translation of this knowledge into predictors for the most relevant, disease specific, Clinical End Points. These tools (imaging biomarkers, models for in-silico medicine research, advanced visualisation tools) will be validated in the application context of two paediatric cancers, Neuroblastoma (NB, the most frequent solid cancer of early childhood) and the Diffuse Intrinsic Pontine Glioma (DIPG, the leading cause of brain tumour-related death in children).

1.1. Scope of the document

This document constitutes the second deliverable of the WP2. It covers all intermediate results of the procedure performed to define the hybrid open data cloud and processing middleware requirements, to finally present the first complete version of the architecture based on these requirements as conclusion result.

The document analyses different key scenarios of the PRIMAGE project (in the form of User Stories) from the point of view the use of data and processing resources which will determine the requirements of the middleware architecture. The domains of these scenarios are basically from the next points of views: i) application managers; ii) application developers; and iii) data managers.

This deliverable will be amended - internally - as the implementation of the ***hybrid open datacloud and processing middleware*** evolves, as new components and needs may arise.

1.2. Target Audience

The document is mainly intended for internal use, although it is publicly released. The main target of this document is the global team of technical experts of the PRIMAGE project, including the experts in; i) the creation of data repositories (WP3); ii) the development of imaging biomarkers (WP4); iii) development of the in-silico models (WP5) in different scales (molecular, tissue and organ); iv) the translation of in-silico models and biomarkers to usable clinical knowledge (WP6) through advanced visualization tools; v) the integration in the PRIMAGE platform (WP7) of all created repositories and developed tools (biomarker, multi-scale models and visualization tools); vi) the execution of prospective studies (WP8) where will be employed the developed tools.

1.3. Structure of the document

This document is organised in nine sections. The second section deals with the description of the procedure employed to analyse the requirements of the **hybrid open datacloud and processing middleware**. After that, the third and fourth sections present the results obtained in the respective phases defined in the procedure ending with the definition of the requirements (section five).

Also, section six discusses the state of the art of the technology that potentially can address the requirements identified. Finally, section seven comprises a proposal for hybrid open datacloud and processing middleware architecture that relies on state-of-the-art components. Section eight presents several sample applications to clarify architecture concepts of the proposal.

9. Conclusion

This document analyses the use cases and requirements that need to be fulfilled by the Hybrid Open Datacloud and processing platform in PRIMAGE. This platform will support the storage and processing in the cloud of the data compiled, curated and enhanced by the medical experts and the data scientists in the project.

In the work of the deliverable we have identified three user roles related to the different interactions we expect with the platform. There will be more user roles in the context of PRIMAGE, but they may access the platform through the high-level applications and services. The roles identified are: Application Developer (who develops applications), Application Manager (who installs applications), Data Manager (who retrieves and stores data). Typically, a full interaction with the platform will imply combining such roles (which can be taken by the same or different people).

The process continued with the identification of 8 user stories that describe in a narrative way the current interactions with the platform and the expectancies of the users from the platform. Such user stories are key to understand the complexity of the process, the processing workflow and the interactions with the different services. From these user stories, we extracted 22 use cases that reflect small-scale interactions with the platform. This use cases will be coded as individual examples for validation and training. The use cases lead to the identification of 32 requirements with respect to security, data storage, infrastructure management and job execution. Those requirements are key to identify a set of technologies that could fulfil them.

The document continues with an analysis of different technologies for security, Storage and Processing that can address the requirements identified. The analysis focuses on state-of-the-art open-source technologies. 15 components have been extensively analysed with respect to the requirements. A table with strengths and weaknesses is presented at the end of each subsection.

With this information, we conclude a proposal of architecture that addresses the use cases. This proposal of architecture does not include a choice in the components, as this will be done after a set of proofs of concept, which will demonstrate the suitability of the use cases. However, a prioritisation of the technologies has been included in the description and we will follow the implementation of the proofs of concept in order.

The architecture identifies the need for the following key services: A Storage Management service able to provide persistent and ubiquitous access to the applications running in the cloud. A Resource Management Service that will configure, scale and deliver fully operational resources for the Job Management services to run both the batch (sequential and parallel) and interactive jobs.

Finally, the document concludes with a description of a set of canonical applications that are representative examples of the applications that are expected in the frame of the project. Therefore, these applications will be used both for the validation of the platform and as examples for the training and development of new applications. Moreover, they can be used for dissemination purposes.