DOI: 10.26314/GARR-Conf18-proceedings-19

The PhenoMeNal Cloud-based Virtual Data Processing and Analysis e-infrastructure

Luca Pireddu[1], Marco Enrico Piras[1], Antonio Rosato[2], Gianluigi Zanetti[1]

[1]Data-Intensive Computing Group, CRS4, [2] Magnetic Resonance Center, C.I.R.M.M.P.

Abstract. PhenoMeNal is a complete software-defined e-infrastructure for data processing and analysis on modern infrastructure-as-a-service (IaaS) cloud platforms, especially tailored for metabolomics. Through the online PhenoMeNal portal, personal instances of the PhenoMeNal Cloud Research Environment (CRE) – which includes widely used data analysis environments (such as Galaxy and Jupyter) and over 250 open source tools – can be easily deployed on-demand, through an automated process, on accessible cloud IaaS like OpenStack, Amazon Web Services or Google Compute Platform, and partner institutional providers such as ReCaS-Bari, Embassy Cloud and de.NBI Cloud. Each PhenoMeNal CRE version fixes the version of the tools it integrates, so deployments are completely reproducible. Thus, PhenoMeNal constitutes a turnkey solution for scientists to easily leverage cloud computing resources to accelerate their work on a user-friendly, workflow-driven, reproducible and shareable data analysis platform.

Keywords. Cloud computing, Infrastructure-as-Code, software containers, metabolomics, reproducibility

Introduction

PhenoMeNal (Peters K., et al. 2018) is a complete software-defined e-infrastructure for data processing and analysis on modern Infrastructure-as-a-Service (IaaS) cloud platforms, especially tailored for metabolomics. This field studies the small molecules in organisms to reveal insights into their metabolic biochemistry. Work in metabolomics typically generates large data sets that require computationally intensive analyses (Sugimoto M., et al. 2012) – e.g., terabytes of data for large cohorts with thousands of metabolite profiles (Joyce A. R., et al. 2006). Cloud computing offers a valuable resource for these computational analyses by allowing access to sufficient computing resources that can be instantiated on-demand and without capital investment.

PhenoMeNal fills the gap between cloud computing and metabolomics researchers by providing a complete and performant data analysis e-infrastructure that includes a large suite of standardised and interoperable metabolomics data processing tools and workflows. The PhenoMeNal e-infrastructure can be easily deployed onto public and private cloud environments, enabling scalable and cost-effective high-performance metabolomics data analysis while hiding the technical complexity from the user. Moreover, PhenoMeNal facilitates reproducible analyses through automated, sharable and citable workflows and through its versioned and reproducible data analysis platform. PhenoMe-

Nal is composed of the Cloud Research Environment, the Portal, and the integrated Scientific Tools and Workflows

1. The Cloud Research Environment

PhenoMeNal's central e-infrastructure component is the Cloud Research Environment (CRE). The CRE can be automatically deployed on most cloud provider services without any sophisticated technical input from the user. Deployments can be activated from the PhenoMeNal web-based portal (described in a later section) or from the command line. The CRE integrates:

- the Galaxy (Afgan E., et al. 2018) and Jupyter (Kluyver T., et al. 2016) data analysis platforms and the Luigi workflow engine;
- over 250 standard metabolomics software tools, packaged by PhenoMeNal;
- ten tested and validated metabolomics data analysis workflows.

Thus, once deployed, the CRE allows the user to start processing data without any further software installation. At the time of this writing, the PhenoMeNal CRE can be deployed on a number of commercial cloud providers, including Amazon Web Services, Microsoft Azure and Google Cloud Platform, and on OpenStack-based private or institutional clouds. Of this latter type, the following partner providers are preconfigured for deployments from the PhenoMeNal portal: EMBASSY Cloud, de.NBI Cloud, and ReCaS-Bari.

From a technical standpoint, the PhenoMeNal CRE is designed as a microservice architecture, using containers to provision microservices for data analysis tools and long-running services such as workflow managers. Kubernetes (k8s) is used to orchestrate containers over the computing infrastructure. PhenoMeNal implements various layers to provision k8s on top of an IaaS. To start, Terraform is used to provision the required computing resources from IaaS – e.g., storage volumes, network resources and compute instances. Terraform provides a cloud-agnostic layer which allows PhenoMeNal to more easily support different providers. The virtual machines are initialised with a customized base image and a Kubernetes cluster with GlusterFS is deployed on them. This entire procedure has been automated and generalized and provided as a new k8s deployment tool called KubeNow (Capuccini M., et al. 2018), which also supports the complete management of the deployment's life-cycle – creation, maintenance, destruction. A KubeNow plugin implements the deployment of the PhenoMeNal-specific services.

This step is performed through the installation of specific Helm charts (the equivalent of a package manager for k8s) which in turn result in the deployment of Docker containers implementing the various services. To realize this solution we created suitable container images for the various services, in addition to the Helm charts themselves; relevant charts and images were contributed to the community – e.g., the Galaxy Helm chart and contributions to the galaxy-kubernetes container images (https://github.com/galaxyproject/galaxy-kubernetes). By default, secure access via https is configured via Cloudflare (http://cloudflare.com), providing dynamic DNS services and encryption for all user communication with services inside the CRE.

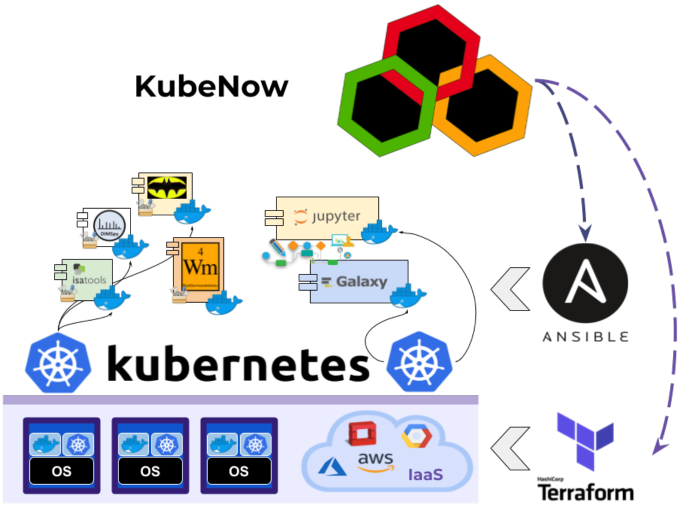
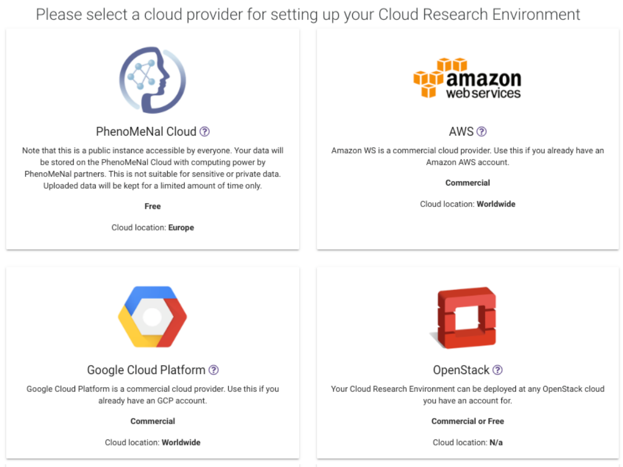


Fig. 1
PhenoMeNal Architecture. Kubernetes is used to provide a uniform platform on any IaaS. All services and tools are packaged as Docker container images and pulled on demand. Terraform and Ansible are used to provision and configure the base infrastructure.

2. The PhenoMeNal Portal

The PhenoMeNal portal (https://portal.phenomenal-h2020.eu) orients users through the features and resources provided by the e-infrastructure and, perhaps most importantly, provides the functionality to deploy and manage CREs through a web interface. Deployments to can be made using an easy-to-follow wizard and managed through a dashboard. The wizard (Fig. 2) collects all required information – e.g., credentials and parameters to access the cloud provider – and then executes the deployment procedure. Interestingly, the user can also customize the computational resources provisioned and the version of the PhenoMeNal to deploy. Once the CRE is operational, the user can follow the links on the dashboard to access the various running services – i.e., Galaxy, Jupyter, etc. – and put them to work, using all the integrated tools and workflows. In addition to this functionality, the portal also gives users access to the PhenoMeNal public instance, which is a public CRE hosted by EMBL-EBI that allows users to test-run a CRE without the need to deploy their own.

Fig. 2
CRE wizard: cloud provider selection during CRE deployment configuration on the PhenoMeNal portal



Please select a cloud provider for setting up your Cloud Research Environment

PhenoMeNal Cloud ⓘ

Note that this is a public instance accessible by everyone. Your data will be stored on the PhenoMeNal Cloud with computing power by PhenoMeNal partners. This is not suitable for sensitive or private data. Uploaded data will be kept for a limited amount of time only.

Free

Cloud location: **Europe**

AWS ⓘ

Amazon WS is a commercial cloud provider. Use this if you already have an Amazon AWS account.

Commercial

Cloud location: **Worldwide**

Google Cloud Platform ⓘ

Google Cloud Platform is a commercial cloud provider. Use this if you already have an GCP account.

Commercial

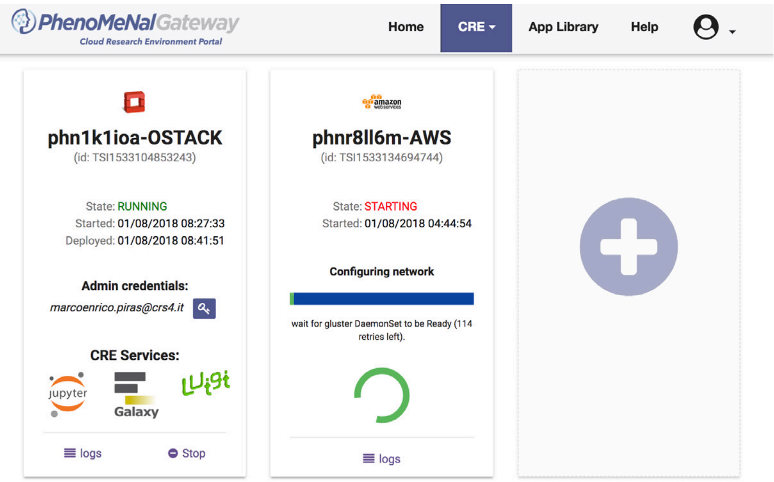
Cloud location: **Worldwide**

OpenStack ⓘ

Your Cloud Research Environment can be deployed at any OpenStack cloud you have an account for.

Commercial or Free

Cloud location: **N/a**



phn1k1ioa-OSTACK
(id: TSI1533104853243)

State: RUNNING
Started: 01/08/2018 08:27:33
Deployed: 01/08/2018 08:41:51

Admin credentials:
marcoenrico.piras@crs4.it

CRE Services:
jupyter Galaxy Luigi

phnr8ll6m-AWS
(id: TSI1533134694744)

State: STARTING
Started: 01/08/2018 04:44:54

Configuring network

wait for gluster DaemonSet to be Ready (114 retries left).

Fig. 3
CRE dashboard on the PhenoMeNal portal.

3. Integrated Scientific Workflows and Tools

PhenoMeNal integrates over 250 internally and externally developed metabolomics tools – which are mainly accessible through the CRE's Galaxy instance. These tools have been containerized by the PhenoMeNal project and integrated in the CRE. As part of the packaging process, we have defined minimal documentation requirements and implemented continuous integration testing. A number of scientific workflows built with these tools are also integrated in the CRE's Galaxy instance; many of these implement analysis procedures from published studies, making it simple to apply them to new data. To keep the CRE deployment relatively light, the container images are dynamically pulled from the PhenoMeNal Docker registry and executed on demand. Furthermore, to support interoperability PhenoMeNal has implemented native support in Galaxy for the standard ISA-Tab and ISA-JSON metabolomics data types (Rocca-Serra P., et al. 2010), coupled with tools to convert between ISA and other formats and to retrieve and store datasets in these formats to MetaboLights (Steinbeck C., et al. 2012) – one of the main metabolomics public data repositories.

4. Extensibility

The PhenoMeNal e-infrastructure can easily be repurposed to other scientific domains by simply changing the tools and workflows that are integrated. Work is already in progress to configure a CRE for genomic data analysis.

5. Conclusions

The PhenoMeNal e-infrastructure allows users to instantiate identical data analysis environments – down to the version of the individual tools – on any compatible IaaS through a simple graphical user interface. Thus, it is a turnkey solution for leveraging the potential of IaaS with very little financial and training investment. Coupled with the possibility of storing, exporting and sharing workflows, PhenoMeNal is a valuable means to ensure the reproducibility of scientific data analysis work. PhenoMeNal can be found at http://phenomenal-h2020.eu. The GitHub repository http://github.com/phnmnl hosts all source code and the Wiki. Source code and documentation are available under the terms of the Apache 2.0 license. Integrated open source tools are available under the respective licensing terms.

6. Acknowledgements

PhenoMeNal is the work of the entire PhenoMeNal project consortium and was funded by the European Commission's Horizon2020 programme, grant agreement n. 654241.

References

Afgan E., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucl. Acids Res., 46(W1), pp W537-W544.

Capuccini M., et al. (2018), KubeNow: an On-Demand Cloud-Agnostic Platform for Microservices-Based Research Environments, arXiv, 1805.06180 (preprint).

Joyce A. R., et al. (2006), The model organism as a system: integrating 'omics' data sets, Nat. Rev. Molecular cell biology, 7(3), p 198.

Kluyver T., et al. (2016), Jupyter Notebooks-a publishing format for reproducible computational workflows, ELPUB, pp 87-90.

Peters K., et al. (2018), PhenoMeNal: Processing and analysis of Metabolomics data in the Cloud, GigaScience (to appear).

Rocca-Serra P., et al. (2010), ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinf., 26(18), pp 2354-2356.

Steinbeck C., et al. (2012), MetaboLights: towards a new COSMOS of metabolomics data management, Metab., 8(5), pp 757-760.

Sugimoto M., et al. (2012), Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis, Curr. Bioinf., 7(1), pp 96-108.

Authors

Luca Pireddu - luca.pireddu@crs4.it

Luca Pireddu coordinates the Distributed Computing Group at CRS4. His current research focuses on novel applications of distributed stream computing and IaaS, with particular interest in problems pertaining to data-intensive biology and smart infrastructures. Luca holds a Ph.D. in Biomedical Engineering from the University of Cagliari, Italy, an M.Sc. Computing Science from the University of Alberta, Canada, and a B.Sc. in Computing Science from Laurentian University, Canada.

Marco Enrico Piras - marcoenrico.piras@crs4.it

Graduated in Computer Engineering at Università “La Sapienza" of Rome, since 2014 he works as a technologist at CRS4. His current research interest deals with the design and implementation of distributed systems for the processing and storage of large volumes of scientific data (Big Data), with a particular focus on the use of microservices architectures and the automation of their deployment.

Antonio Rosato - rosato@cerm.unifi.it

Antonio Rosato is Associate Professor of Chemistry at the University of Florence. He holds a Ph.D. in Chemical Sciences from the same University. His research interests span from structural biology of metalloproteins to NMR methods in the life sciences, and all associated computational aspects. He has co-authored more than 100 articles in scientific journals.

Gianluigi Zanetti - gianluigi.zanetti@crs4.it
Gianluigi Zanetti is the director of the Data-Intensive Computing sector at CRS4. He holds a degree in Physics from the University of Bologna and a Ph. D. in Physics from The University of Chicago. His research spans many areas of computational science and is widely published in major journals and conferences (over 100 refereed publications).