EGI-Engage

Evaluated cloud environment and demonstrator of analysis workflow for biobank studies

D6.16

Date	18 July 2017
Activity	WP6
Lead Partner	BBMRI-ERIC
Document Status	FINAL
Document Link	https://documents.egi.eu/document/3020

Abstract

This deliverable summarizes the three technology demonstrators developed within BBMRI Competence Centre in EGI-Engage: (a) proteomic workflows deployed by BBMRI.cz, (b) complex genomic workflows deployed by BBMRI-NL, (c) BiobankCloud-based workflows deployed by BBMRI.se. All of the tools have been made available open source and the workflows have been validated not only on EGI.eu infrastructure, but also inside the private clouds of the biobanks to allow for processing of very sensitive personal data.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Boris Parak, Cuong Duong Tuan, Ondrej Vojtisek, Roan Kanninga, Gerben van der Vries, Pieter Neerincx, Enol Fernández, Morris Swertz, Altti Ilari Maarala, Davit Bzhalava, Jim Dowling, Petr Holub	BBMRI.cz, BBMRI.nl, KTH, KI, EGI.eu, BBMRI-ERIC	2017-09-09
Moderated by:	Małgorzata Krakowian	EGI Foundation	
Reviewed by	Zdenka Dudová Heimo Müller	Institute of Computer Science, Masaryk University Medizinische Universität Graz	2017-09-13 2017-09-18
Approved by:	AMB and PMB		2017-09-27

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	2017-08-30	Initial version with BBMRI.cz and BBMRI.se.	*Boris Parak, Cuong Duong Tuan, Ondrej Vojtisek, Enol Fernández, Altti Ilari Maarala, Davit Bzhalava, Jim Dowling, Petr Holub*
v.2	2017-09-09	Added BBMRI-NL and reviewed Intro/Conclusions/Excecutive Summary.	*Roan Kanninga, Gerben van der Vries, Pieter Neerincx, Morris Swertz, Petr Holub*

TERMINOLOGY

A complete project glossary and acronyms are provided at the following pages:

- https://wiki.egi.eu/wiki/Glossary
- https://wiki.egi.eu/wiki/Acronyms

Contents

Executive summary

Medical research dealing with human data is very much influenced and constrained by the data protection regimes and necessary trust that medical research must maintain with its research participants. This makes straightforward implementation of scalable cloud computing very complicated. On the other hand medical research is facing big data storage and processing challenges often due to advent of new technologies, such as omics data generation, and rapidly decreasing costs of data generation.

BBMRI Competence Centre has implemented 3 demonstrators summarized in this document, primarily aiming at deployment of the cloud computing within the private clouds of the biobanks in order to comply with the data protection regulations, based on the EGI.eu cloud stack. These range from (a) relatively simple implementation of proteomics pipeline based on orchestrating existing distributed operation of an existing commonly used software (implemented in BBMRI.cz), to (b) complex genomics pipelines using common bioinformatics workflows (implemented in BBMRI-NL), to (c) dedicated development of metagenomic workflows for BiobankCloud platform, which enables advanced security features and multi-tenancy aware access control (implemented in BBMRI.se). The workflows have been tested with the medical researchers, mostly inside the biobanks (except BBMRI-NL), and are aimed to be demonstrated as a part of IT tools of BBMRI-ERIC during Global Biobank Week 2017.

The results of the demonstrators show the technical feasibility of both operating on EGI.eu infrastructure as well as in private clouds of the biobanks using the EGI.eu software stack, when sensitive personal data is processed (either without informed consent for diagnostic/therapeutic purposes, or with informed consent or legal environment not allowing processing in public clouds). The work done in the BBMRI Competence Centre also reveals urgent need for further continuing the work on defining common Europe-wide policies to enable ingesting third-party cloud resources into the logical private space of the biobanks to allow scaling up the computations even further - possibly within European Open Science Cloud.

1 Introduction

1.1 Background and motivations

The BBMRI Competence Centre (BBMRI CC) in EGI-Engage aims at developing the workflows that are typical for biobanks storing human biological samples and associated data. Therefore the demonstrators developed in the BBMRI CC involve sensitive personal data - personal data of research participants (patients or donors), either for research or for diagnosis/treatment purposes. The biobanks are facing a number of computationally-heavy and storage-heavy tasks recently, in particular with the advent of various data-intensive omics technologies - genomics, proteomics, metabolomics, etc. Beyond the technical aspects of effectively scalable processing, the limiting factors for implementing these workflows is informed consent provided by the research participant (maybe missing for diagnosis/treatment purposes of individual patients - this is still relevant for the workflows relevant to biobanks being part of the clinical process) and by the European and national legal regulations related to data protection, good research practice and medical research in general. The aim of the BBMRI CC was to attempt to pilot these workflows in order to demonstrate utility of scalable cloud computing to the biobanking community to raise their interest in efficiency improvements of their current, often poorly performing workflows. It is also important to notice that medical research involves often researchers with very limited IT background and thus simplicity of use of the workflows is a very important parameter. Validation of the demonstrators is targeted to (a) demonstrate technical parallelization and distribution of the computations, (b) demonstrate the computations using real sensitive data from the biobanks. Separation of those two steps allows to do at least partial validation in case that the deployment within the biobanks become problematic for any reason.

The BBMRI CC has already delivered a number of deliverables that are relevant for this description of demonstrators:

- General privacy and data protection requirements have been outlined in EGI-Engage M6.2 document.
- The description of the workflows relevant for implementing the demonstrators has been provided in the EGI-Engage Deliverable D6.8.
- EGI-Engage is supported, to resource-limited extent, further development of the BiobankCloud platform (developed in BiobankCloud project, 106495) in order to make it compatible with EGU.eu platform stack (enabling OCCI interfaces) and integrating it with the federated SAML authentication (based on Shibboleth, but also enabling BBMRI AAI, EGI CheckIn, or future Lifesciences AAI developed as a part of CORBEL cross-life-sciences Research Infrastructure platform). This has been described in the EGI Deliverable D6.11.

While the BiobankCloud is almost ideal solution for the purposes of sensitive human data due to its advanced security features and multi-tenancy aware access control, it has also showed relatively high barrier for adoption due to the required workflow paradigms. There have been two demonstrators have been successfully developed: proteomic workflows and genomic workflows. While both have been implemented using common, relatively simple cloud technologies available within the EGI.eu infrastructure, the genomic workflows in particular have been also implemented using the BiobankCloud platform.

1.2 Infrastructure

EGI.eu Infrastructure has been successfully used for developing and piloting both the proteomic and genomic workflows on two EGI.eu providers: CESNET-Metacloud in Czech Republic and BELNet in Belgium. However, in most cases, unless very liberal informed consent is provided by the research participants, the infrastructure provided by third parties is very hard to be used by the biobanks. Hence the default operating mode for the time being is to use the private clouds that are often being set up by the biobanks internally in the last several years. As a part of the demonstrators, we have therefore at least utilized the middleware stack provided by EGI.eu to deploy private cloud within the biobanks as a default option (demonstrated, e.g., using proteomic workflows within MMCI biobank as a part of BBMRI.cz). Within the BBMRI CC we have also looked into the certification (e.g., ISO 27001+27017+27018) as a possible enabler for the biobanks to enhance scalability of their private clouds using third-party provided cloud resources. This process has however turned out to be extremely difficult partially also due to changing legal landscape (EU General Data Protection Regulation replacing Directive 95/46/EC with yet uncertain ramifications, heterogeneous implementations into the national laws and ongoing developments of relevant Codes of Conduct), and thus BBMRI CC achieved only very preliminary steps and more work is needed, ideally within the scope of the European Open Science Cloud.

1.3 Proteomics Workflow Demonstrator (BBMRI.cz)

1.4 Overview

Name	*Deployment of Windows-based graphical application for processing of mass-spectrometry proteomics data to cloud environment*
URL	*N.A.*
Description	*The SW tools used to analyse proteins are often Windows-based with GUI frontends for interactive visualization of the processing, which makes them uneasy deployable to cloud environment. Our pilot presents a way how to deploy Skyline, which is widely used in proteomics. Communication with it is maintained using queue, messages and command line interface (CLI) without direct access of user to the workstation.*
Value proposition	*The computational capacity required by the processing is the limiting step in generation of data from the mass spectrometry devices, substantially limiting the overall throughput of those. The requirements for the analytic SW: fast, accessible to multiple users and able to solve multiple tasks at once. It is costly to provide a powerful set of workstations to everyone in a laboratory, and it needs additional coordination or SW tools to maintain access of multiple users to a single resource. Scalable cloud resources also allow to better respond to peaks in required processing capacity, e.g., when a batch of samples from patients arrives and the results are needed fast for treatment/therapeutic purposes. EGI middleware allows for building these clouds+applications also inside the hospitals.* *The designed concept enables user to have the analytic environment exclusively for whenever he/she wants.*
Customer/user of the demonstrator	*Pilot is primarily targeting researchers dealing with patients data under informed consent (typical scenario for BBMRI-ERIC and its partner biobanks). Secondary, but also relevant, is biobanks analysing the data for diagnostic/therapeutic purposes (without informed consent).*
Scenario	*Two scenario have been tested: 1) deployment on MetaCentrum infrastructure to validate technical feasibility and assess scalability (insensitive data - such as calibration samples or data with informed consent from patients where such processing is explicitly allowed), 2) deploy internally in hospital for real patients' data (sensitive data with or without informed consent for research)*
Success criteria	*User (researcher) is able to analyze output data from mass*

	spectrometer in SW Skyline in a scalable way (but theoretically any windows based SW with CLI).
User Documentation	*https://github.com/cduongt/skyline-rabbitmq*
Technical Documentation	*https://github.com/cduongt/skyline-rabbitmq*
Developer team	*Boris Parak, Cuong Duong Tuan, Ondrej Vojtisek*
License	*MIT. Licenses of used tools are described on https://github.com/cduongt/skyline-rabbitmq*
Source code	*https://github.com/cduongt/skyline-rabbitmq*

1.5 Architecture

The architecture consists of three parts: user workstation(s), master node and worker node(s). Workstation is a node where the data is stored and a user wants to analyse it. It is common office computer or laptop without any specialized HW. Master node is a virtual server (Linux) that provides Samba storage to enable exchange of data among workstations and worker nodes and that manages computational tasks of worker nodes. Worker nodes are virtual machines running Windows with Skyline installed. When the user submits a task (by copying the files into shared samba folder), the master node assigns the task to one of the worker nodes. Communication between master and worker is implemented with RabbitMQ (open-source message broker). The task is triggered (on master node) by a Python script that controls the shared folder. When new “filename.sky” file is uploaded, the master node allocates a task to the worker node and creates “filename.progress” to label that this task is being processed. There is a running Python script, which controls the communication also from the side of the worker node. Script accepts a message from the master node and starts processing data in the Skyline application. The application is started using SkylineRunner.exe, which is an executable launching Skyline without user interface and is controlled using command line interface. When the data are processed the output (.skyd file) is stored inside shared samba folder.

Information about the process are stored into .log files located on the shared file system.

The same approach can be reused theoretically for any SW which runs on Windows and has CLI (e.g. MaxQuant).

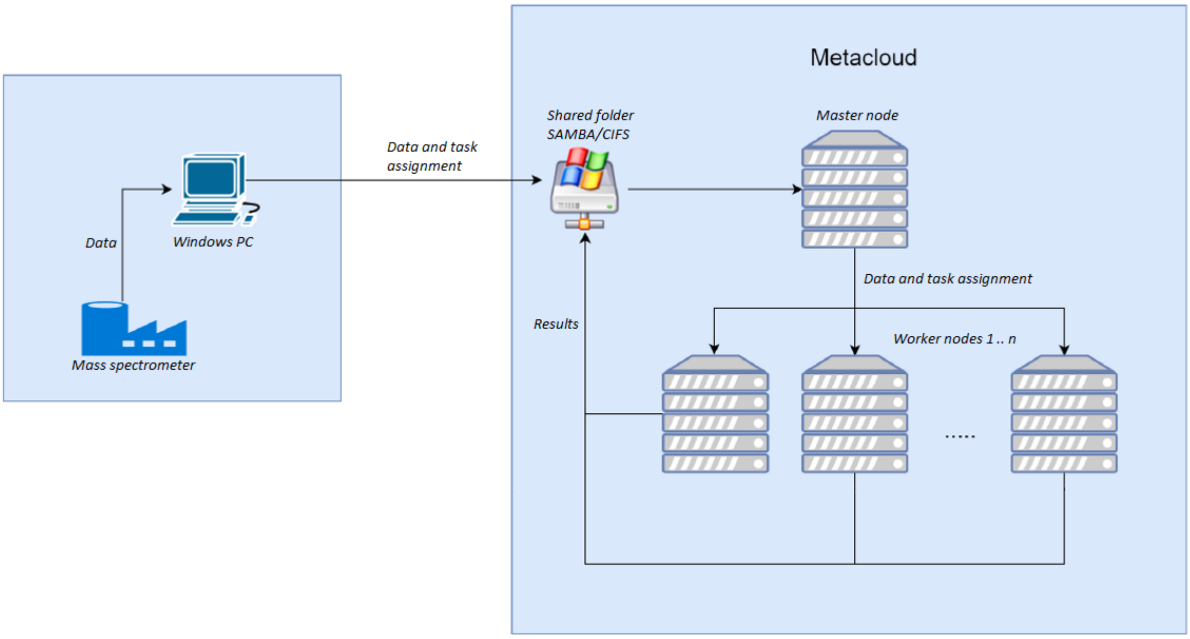


Figure 1: Skyline pipeline overview.

1.6 Demonstration

1.6.1 Scenario

Two scenarios have been tested. First was deployment of architecture to “public” cloud (CESNET-MetaCloud) which represents a situation of insensitive data. The second scenario represents a situation when data is sensitive (due to the common forms of informed consent and legal restrictions in some countries, and in some cases even without informed consent for diagnostic/therapeutic purposes) and can’t leave the institution (hospital and/or laboratory). In this case the existing VMware infrastructure in hospital has been utilized. Public demonstrator of the technology is scheduled for the Global Biobank Week 2017 (www.globalbiobankweek.org) as a part of the IT tools demonstrated by BBMRI-ERIC.

In both scenarios the end-user maps network drive (shared Samba folder). Then he/she prepares task in Skyline at his/her workstation. Task in this context means a combination of .sky file and related raw measured data. At this point user can analyse the data directly at the workstation or in cloud using the described tool. Difficulty of the analysis depends on the raw data volume which is analysed. The efficiency is affected by a HW used at a server site and by a network capacity among workstation, master node and worker node. As it turned out during evaluation of this demonstrator in a third party cloud infrastructure (MetaCloud in this case), a slow network (approximately 10 MB/s) between the laboratory and the cloud provider can be a significant bottleneck.

For the testing purposes we were provided by sample data (combination of .raw and .sky files) containing measurements of calibration samples.

1.6.2 Feedback from the user

The demonstrator has been evaluated with the laboratory staffs, who primarily involve domain experts and not IT experts. The collected feedback can be summarized as follows:

The tool is missing more interactive way for user to control his/her tasks. User should be able to execute tasks manually whenever wants and he/she should be better (.log files considered inconvenient) informed about the process of current task. Dynamic creation of virtual machines and improvement of the *.raw files definition process belong to the given task and are needed to be scalable with ability to solve batch tasks.

1.7 Future plans

The implemented and tested tool is only a proof of concept and further deployment will be necessary:

- Worker nodes should be created (and erased) dynamically on purpose.
- Parse data from .sky file so the paths are not hardcoded.
- More interactive way of task execution - start, stop, process information, …
- Optimization.

The connection between user and shared file system can be significant bottleneck.

Scaling up of the infrastructure available for the biobanks to cover their computational and storage needs further depends on creating common guidelines for processing privacy sensitive information using third-party provided cloud resources into the logically private cloud of the biobank. As this is related to all the demonstrators in this Deliverable, further discussion follows in the common section on future work as a part of Conclusions.

2 Genomics Workflow Demonstrator (BBMRI.nl)

2.1 Overview

Name	*NGS_DNA analysis pipeline - for whole genome sequence data processing using virtual SLURM cluster*
URL	*NA*
Description	*Combination of Ansible playbooks to deploy a complete SLURM cluster on EGI cloud and DNA analysis pipeline including all dependencies and reference data.*
Value proposition	*In research context BBMRI-NL partner regularly have to analyse large cohorts of whole genome DNA data (next generation sequencing, NGS, 100s of samples, 50+ TBs of data) in the context of large multi-center studies (requiring 20+ nodes of 24-core/256Gb RAM and 100TB shared storage). However, the capacity for these large scale analyses is typically not available in the partner research institutes greatly impeding the analyses.*
Customer/user of the demonstrator	*We delivered a portable pipeline implementation of industry standard DNA NGS processing steps (BASH) deployed using Ansible, EasyBuild, and executed using module system for binaries and SLURM as resource manager for parallel processing which can be run on EGI cloud.*
Scenario	*This pilot demonstrates how to farm-out large DNA analyses using public data (exome) on a demonstration cluster of 4 worker nodes, 8 cores each.*
Success criteria	*Primary goal was to evaluate ease of deployment ideally using single Ansible command (completed; we tested on EGI cloud cluster as well as on independent cluster). Secondary goal was to evaluate if the system does not have performance bottlenecks preventing scaling out of these analyses (first experiments successful; further experiments are needed to assess scaling up).*
User Documentation	*https://molgenis.gitbooks.io/ngs_dna/*
Technical Documentation	
Developer team	*Roan Kanninga, Gerben van der Vries, Pieter Neerincx, Enol Fernández, Morris Swertz*
License	MIT
Source code	pipeline: https://github.com/molgenis/NGS_DNA playbook: https://github.com/bbmri-nl/bbmri-nl-pipeline-deployment

2.2 Architecture

The NGS_DNA analysis pipeline consists of a series of command-line tools that need to run in a specific order as jobs.

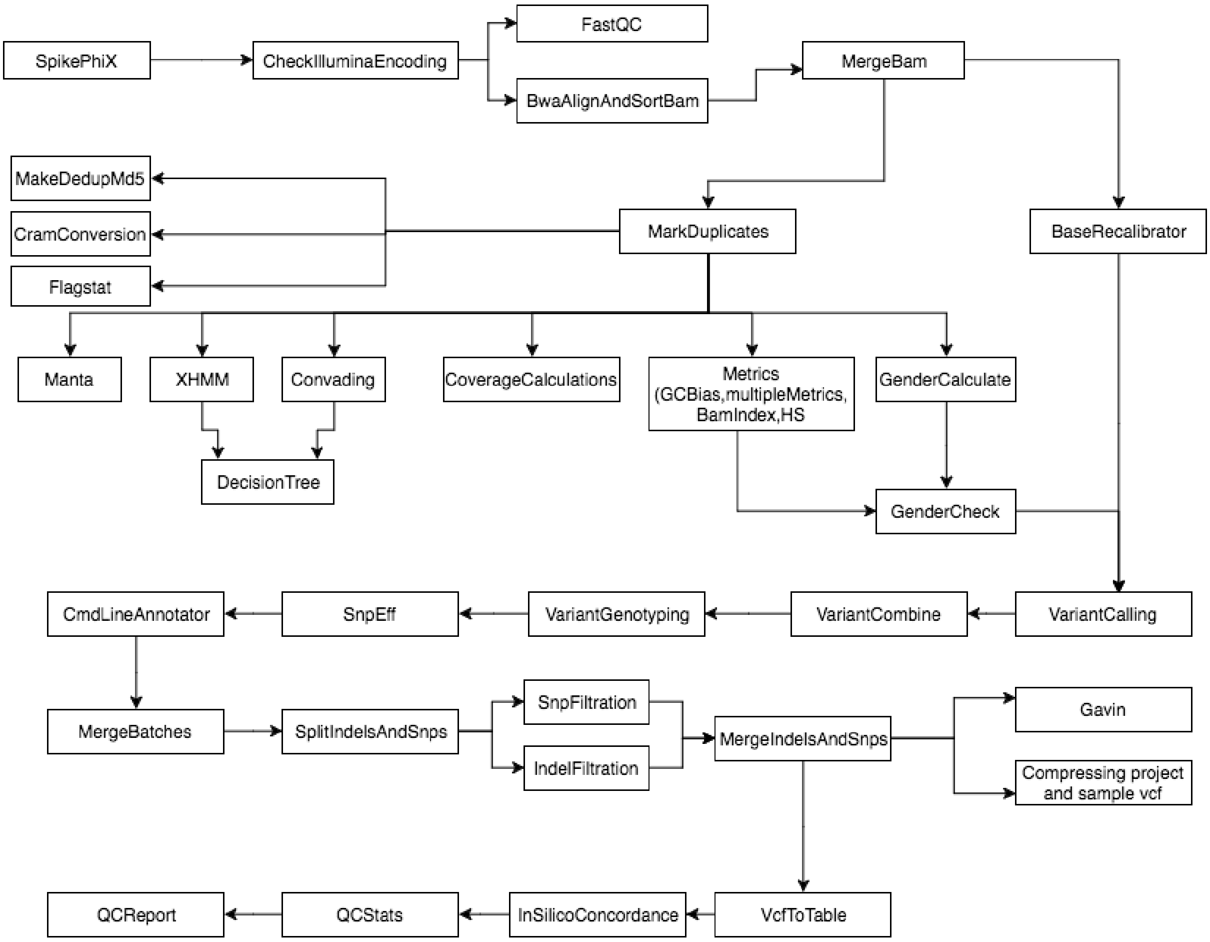


Figure 2: NGS pipeline overview.

We used a subset of the standard NGS pipeline (which is available complete online) for testing. The pipeline steps are defined using the MOLGENIS open source compute system[1], which can automatically generate the analysis jobs as defined in the manual. These jobs are managed by a batch scheduler such as SLURM and require a shared storage with data and tools that can be accessed by all the nodes involved in the computation. For the demonstrator, we have deployed

[1] H.V. Byelas, M. Dijkstra, P.B.T. Neerincx, F. van Dijk, A. Kanterakis, P. Deelen, M.A. Swertz (2013) Scaling bio-analyses from computational clusters to grids. Proceedings of the 5th International Workshop on Science Gateways (IWSG 2013)

virtual SLURM clusters on top of EGI FedCloud infrastructure using Ansible playbooks that enables re-building clusters as needed on new resources.

The virtual clusters are first created using the AppDB VMOps graphical dashboard that allows authorized users to manage VMs on the FedCloud. VMOps organises VMs in topologies, each topology being a set of VMs with the same specifications. For our case, each cluster is composed by two different topologies:

- A head node with the SLURM controller and two volumes attached one for home directory and one for general data storage. Both volumes are persistent and exported via Network File System (NFS) to the worker nodes.
- A set of SLURM workers that execute the jobs.

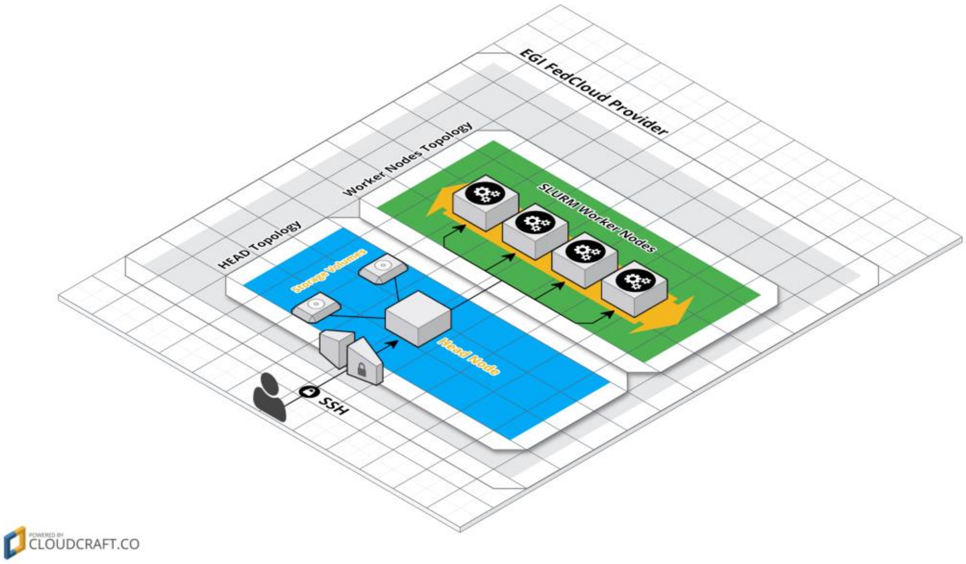


Figure 3: SLURM Virtual Cluster on EGI FedCloud Provider

Once the topologies are running on the cloud providers, an Ansible playbook is used to setup the environment where the pipelines can be further deployed by users. This playbook configures the following items:

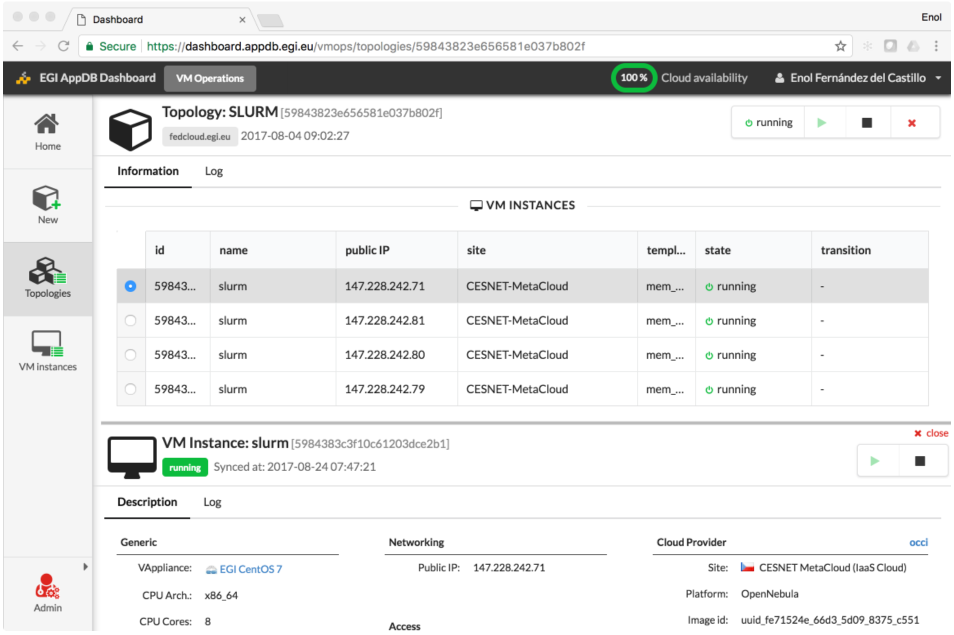
- Sets up the hostnames of each node so they are consistent with the SLURM configuration.
- Sets up the firewall to only allow connections between the nodes of the topologies and ssh from any external hosts.
- Set up NFS server at the front node to export the available volumes and mount those volumes at the worker nodes.
- Set up SLURM at the front node and worker nodes.

This playbook is run from the head node, uses existing roles developed by INDIGO-DataCloud, and is available at Ansible Galaxy[2]. Some bugs found in the roles are being discussed with the original developers for including them upstream.

Virtual clusters can be easily cloned on new providers in few minutes by repeating the process of creating the topologies on VMOps dashboard and running the playbook on the new nodes. The Worker nodes topology can also be scaled up/down as needed for adapting to each pipeline execution. The VMs are only accessible by the user creating them in the AppDB VMOps and are protected by the external firewall of the resource provider and by an internal firewall that only allows communication within the cluster itself and ssh access from any other host within the same resource provider or external to it. Login accounts are only accessible via SSH keys and passwords are not allowed in order to avoid potential brute force attacks. As with the VMs, the storage volumes are only accessible by the user creating them and exported via NFS only to the nodes of the cluster.

Currently there are two deployed clusters on two EGI providers: CESNET-Metacloud in the Czech Republic and BELNet in Belgium. Both are configured identically with:

- 1 head node with 2 cores and 2GB RAM running CentOS 7 with a 20 GB volume for home directories and 4TB for general use. Both volumes exported via NFS.
- 4 worker nodes with 8 cores and 8GB RAM each running CentOS 7 that mount the NFS-exported volumes and run the SLURM jobs. Figure below shows the topology for the worker nodes in CESNET-Metacloud as configured in AppDB VMOps.



Topology: SLURM [59843823e656581e037b802f]
fedcloud.egi.eu 2017-08-04 09:02:27
running

Information Log

VM INSTANCES

	id	name	public IP	site	templ...	state	transition
●	59843...	slurm	147.228.242.71	CESNET-MetaCloud	mem_...	running	-
○	59843...	slurm	147.228.242.81	CESNET-MetaCloud	mem_...	running	-
○	59843...	slurm	147.228.242.80	CESNET-MetaCloud	mem_...	running	-
○	59843...	slurm	147.228.242.79	CESNET-MetaCloud	mem_...	running	-

VM Instance: slurm [5984383c3f10c61203dce2b1]
running Synced at: 2017-08-24 07:47:21
close

Description Log

Generic
VAppliance: EGI CentOS 7
CPU Arch.: x86_64
CPU Cores: 8

Networking
Public IP: 147.228.242.71

Access

Cloud Provider occi
Site: CESNET MetaCloud (IaaS Cloud)
Platform: OpenNebula
Image id: uuid_fe71524e_66d3_5d09_8375_c551

Figure 4: AppDB VMOps Dashboard screenshot with details of the SLURM worker nodes topology at CESNET MetaCloud center of EGI FedCloud

[2] https://galaxy.ansible.com/grycap/slurm/

Once the virtual Slurm cluster has been deployed a second Ansible playbook was used to deploy:

- Create a basic folder structure for modules, software, user data, etc.
- Lmod (Lua implementation of an Environment Module System - https://github.com/TACC/Lmod).
- EasyBuild (software build and installation framework - https://github.com/easybuilders/easybuild).
- All bioinformatics software using EasyConfigs and the EasyBuild framework.
- Reference data (like the human reference genome sequence).

The Environment Module System is used to resolve dependencies at runtime. The EasyBuild framework is used to install the bioinformatics software without the need of root privileges. A key feature of the EasyBuild framework is that it easily captures customizations in non-default installation procedures for software that is not decently packaged in a recipe.

On your own machine you run the Ansible playbook. This playbook does all the work for you when provided your SSH keys. The playbook will access the EGI cloud cluster head node, create the folder structure, and installs all the software and reference data needed. Then the cluster is ready to use. To run an analysis you login as 'slurm user', upload your sample data and run the pipeline as described in the standard MOLGENIS compute NGS DNA pipeline manual.

Full stack of necessary binaries is described at

https://molgenis.gitbooks.io/molgenis-pipelines/content/pipelines/ngs-dependencies.html

2.3 Demonstration

2.3.1 Scenario

To evaluate the pipeline deployment into EGI cloud we deployed the pipeline together with the sequence data of a publicly standardized test sample (NA12878). We used only core analysis dependencies and skipped optional downstream steps to speed up the demo and make it more comprehensible.

We evaluated the pipeline within the UMC Groningen diagnostic bioinformatics team, from deployment until running of an analysis using the test sample. We also plan to demonstrate these results both in the BBMRI-NL national compute and pipeline working group meeting, as well as during Global Biobank Week in September 2017. Then we will demonstrate the deployment procedure as well is the starting of the pipeline.

2.3.2 Feedback

The bioinformaticians who evaluated the pipeline reported that the running of the Ansible playbook was trivial, and that they would like to use the same technology for their other clusters, when processing personal data (i.e., the same scenario as deploying private clouds inside the biobank). They said that it was easier than they had expected. Only criticism was that the default cluster configuration is still a bit barebone, e.g. you still need to install basic tools like 'nano'. What was also appreciated is that you can start your analysis with a completely empty cluster which makes it easier to keep the analysis reproducible. The NGS_DNA pipeline using MOLGENIS compute is easy to configure using a list of samples.

2.4 Future plans

To ensure reproducible analyses the Ansible playbook method is perfect and we plan to expand the work method to all analyses. EGI cloud is also promising because it enables us to scale up/down very quickly. This we certainly want to study further. However, before we can implement EGI cloud at large scale we must get assurances about data security in light of human subjects' samples having privacy issues. This was an issue which was found earlier in the BBMRI Competence Centre in EGI-Engage, but will take more time to converge to a solution accepted Europe-wide, as further discussed in the Conclusions of this Deliverable. Ideally we would integrate with ELIXIR/BBMRI developments towards federated authentication/authorization, which is now under development in CORBEL Project as LifeSciences AAI. In addition the cost structure must be fleshed out. We look to local, national and EU computes institutes to take the lead here.

3 BiobankCloud Workflows (BBMRI.se)

3.1 Overview

Name	*BiobankCloud NGS Workflow - ViraPipe: Scalable Parallel Pipeline for Viral Metagenome Analysis from Next Generation Sequencing Reads*
URL	*NA*
Description	*This is a technology demonstrator of the pilot NGS analytical software developed for the BiobankCloud secure high-performance platform.*
Value proposition	*The BiobankCloud-based NGS analysis demonstrates setup of secure multi-tenant aware cloud service. While his early prototype has demonstrated high scalability using sensitive human data, at cost of developing custom pipeline components for BiobankCloud (using underlying Apache Spark).*

Customer/user of the demonstrator	*Researchers working in the field of metagenomics. In the broader sense it shows also to IT and bioinformatic software developers how the similar pipelines can be implemented for the secure multi-tenant BiobankCloud environment.*
Scenario	*Metagenome analysis using 768 human samples, deployed at the BiobankCloud cluster at Karolinska Instituet.*
Success criteria	*Users from biological or medical research are able to utilize the BiobankCloud platform for metagenome analyses.*
User Documentation	https://github.com/NGSeq/ViraPipe
Technical Documentation	https://github.com/NGSeq/ViraPipe
Developer team	*Altti Ilari Maarala, Davit Bzhalava*
License	*MIT License*
Source code	https://github.com/NGSeq/ViraPipe

3.2 Architecture

Next Generation Sequencing (NGS) technology enables identification of viral genomes directly from microbial communities more rapidly and cheaper than ever before. However, traditional computational algorithms, data formats and pipelines for metagenome analysis are developed for sequential processing whereas technology is moving towards parallel and distributed computation. Thus, there is an urgent need for metagenome analysis pipelines to utilize all the power of data parallel computation. ViraPipe is a massively scalable viral metagenome analysis pipeline capable to analyse thousands of individual human NGS samples in parallel in tolerable time.

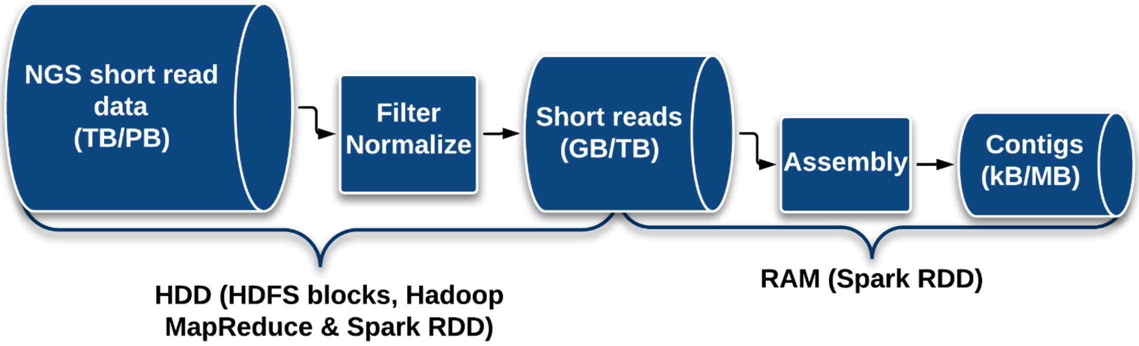

NGS short read data (TB/PB)
Filter Normalize
Short reads (GB/TB)
Assembly
Contigs (kB/MB)
HDD (HDFS blocks, Hadoop MapReduce & Spark RDD)
RAM (Spark RDD)

Figure 5: ViraPipe metagenome analysis pipeline overview.

ViraPipe is Apache Spark based pipeline for analysing viral metagenomes from NGS read data in a computing cluster. The pipeline is designed especially for identifying viral genomes, but the software is applicable for any other large-scale genome analysis purposes. The pipeline integrates parallel BWA read aligner, MegaHit DeNovo assembler, and BLAST and HMMER3 sequence search tools. ViraPipe was deployed and evaluated on the BiobankCloud/Hopsworks platform.

3.3 Demonstration

3.3.1 Scenario

We show the scalability of ViraPipe by running experiments on mining virus related genomes from metagenomic datasets in a distributed Spark computing cluster. The results show linear scalability of the pipeline and ViraPipe analyses 768 human samples in 363 minutes on a Spark computing cluster comprising 20 worker nodes and 1120 cores in total.

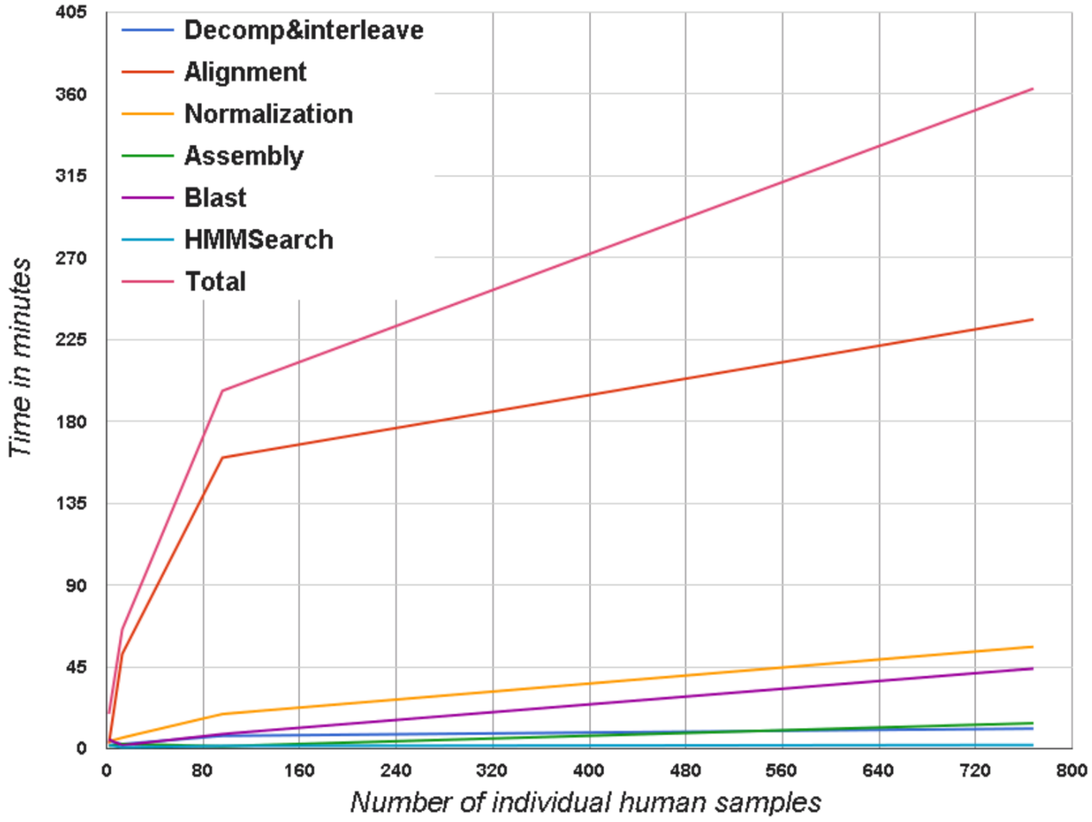


Figure 6: Scalability evaluation of the ViraPipe pipeline.

From the results above, it can be seen that total running time increases superlinearly when the number of samples is small (< 100) and converges to linear growth when the number of samples increases.

3.3.2 Feedback

Karolinska University and Aalto University developed ViraPipe and they were impressed with the development and deployment environment of BiobankCloud/Hopsworks. The platform enabled the deployment of a wide range of software tools and their integration in a single Spark-based pipeline.

3.4 Future plans

BiobankCloud has been demonstrated a very scalable secure computing platform with multi-tenancy aware access control. The downside is that only a few bioinformatic tools support the computing paradigms used by BiobankCloud and need to be implemented anew, as demonstrated by the in this demonstrator. The anticipated future work to be implemented as a part of collaboration between BBMRI.se and KTH is to port other types of computations to the BiobankCloud, to make it more readily available for researchers coming from biology or medical domains without software development expertise. BiobankCloud is provided by Karolinska Instituet as a Platform as a Service as a part of the BBMRI.se services.

4 Conclusions, lessons learnt, exploitation of results and future work

As discussed in this document, medical research dealing with human data is very much influenced and constrained by the data protection regimes and necessary trust that medical research must maintain with its research participants. This makes straightforward implementation of scalable cloud computing extremely complicated. On the other hand medical research is facing big data storage and processing challenges often due to advent of new technologies, such as omics data generation, and rapidly decreasing costs of data generation. Furthermore, medical research has tremendous potential regarding societal impact of the research, thus rendering attempts to push the boundaries of privacy-respecting scalable cloud computing very important.

BBMRI Competence Centre has implemented 3 demonstrators summarized in this document, primarily aiming at deployment of the cloud computing within the private clouds of the biobanks in order to comply with the data protection regulations. These range from relatively simple implementation of proteomics pipeline based on orchestrating existing distributed operation of an existing commonly used software, to complex genomics pipelines using common bioinformatics workflows, to dedicated development of metagenomic workflows for BiobankCloud platform, which enables advanced security features and multi-tenancy aware access control. The workflows have been validated with the medical researchers inside the biobanks and are aimed to be demonstrated as a part of IT tools of BBMRI-ERIC during Global Biobank Week 2017[3].

Common lessons learned.

- **For biobanks:** The developed technology demonstrators show feasibility of scaling up both simple and complex workflows, from technical perspective in any type of clouds: private or public or hybrid. Given the legal and organizational uncertainties using third-party provided resources, it is always possible to build private clouds inside the biobanks and use the workflows such as the ones demonstrated in this deliverable. Biobanks have been given clear information on the licensing of the workflows and its community-backed open-source is an efficient approach to sustainability.
- **For bioinformatic developers and EGI.eu:** Workflows developed for the medical researchers often need to be very approachable from users' perspective and one cannot assume ability to develop custom software or even to implement significant customizations of the workflows.

[3] Parák, Boris, Tuan, Cuong Duong, Vojtíšek, Ondřej, Kanninga, Roan, van der Vries, Gerben, Neerincx, Pieter, … Holub, Petr. (2017). Cloud Workflows Developed in BBMRI Competence Center of EGI-Engage. Zenodo. http://doi.org/10.5281/zenodo.898189

- **For EGI.eu:** Biobank users often require software depending on particular operating systems and/or licensing (e.g., Windows for proteomics workflows), making highly scalable computing dependent on additional variables.
- **For biobanks and EGI.eu:** Further work is needed on settling on consensus for requirements/guidelines on ingesting third party cloud resources into the logical perimeter of private clouds within biobanks. This is dependent on national and international regulatory frameworks and will require very substantial long-term investment to cope with the ever-changing legal/ethical/societal landscape - making it a very good candidate for target of the upcoming European Open Science Cloud.

Future work on using clouds for privacy-sensitive data. All the three demonstrators described in this Deliverable would substantially benefit from Europe-wide guidelines for using the cloud infrastructures provided by third-party organizations. Smaller to medium sized BBMRI-ERIC biobanks have often very limited cloud resources available internally, if any at all. Hence being able to use third-party provided resources, e.g., by logically ingesting them into their private clouds under the same or better data security guarantees as for their own resources, would allow the biobanks to scale up their infrastructure and in particular to meet peak needs for computing capacity. The use of the third party resources needs to be compliant to the upcoming European General Data Protection Regulation (GDPR) and its national implementations. The community will hopefully benefit here from the Code of Conduct for Use of Medical Data in Research based on GDPR Article 40, which is currently under development coordinated by BBMRI-ERIC and including many of the leading research organizations dealing with medical/health data as well as various relevant advocacy groups. Within the BBMRI Competence Centre of EGI-Engage we have explored use of ISO 27018 certification, but the work has not been resulted in any specific recommendation due to lack of evidence of its successful use for other similar domains, and pioneering it being outside of the resources and time frame given to the Competence Centre. BBMRI-ERIC has however long-term commitment to developing a solution for this issue, and hence further work has been become a topic for ongoing and upcoming projects related to European Open Science Cloud (EOSC). BBMRI-ERIC is hence responsible for preliminary development of those policies as a part of EOSCpilot (WP3, Task 3.1.4.) and EOSC-hub projects (WP2, Task T2.4). Full development and wide adoption of these policies will however require a substantial funding and very likely a separate project within the EOSC framework.

Exploitation of results as a part of BBMRI-ERIC infrastructure. BBMRI-ERIC is planning to integrate the developed proteomics and genomics workflows into the open-source reference tools BIBBOX[4], which is developed and maintained as a part of BBMRI-ERIC infrastructure in Common Service IT[5]. BIBBOX builds on virtualization to deliver ready-to-use and ready-for-integration components to

[4] http://bibbox.bbmri-eric.eu/

[5] http://www.bbmri-eric.eu/BBMRI-ERIC/common-service-it/

build IT infrastructure inside the biobanks, ranging from biobank and laboratory information systems to data processing tools. Hence the developed workflows fully fit into this system. BiobankCloud will be further run at Karolinska Instituet as a part of infrastructure provided by BBMRI.se, which is in principle accessible to researchers Europe-wide. The proteomic workflow has been also suggested for PhenoMeNal metabolomics infrastructure[6], as a possible component.

[6] http://phenomenal-h2020.eu/