Outcome summary

PhenoMeNal IT security discussion with NeIC/Tryggve2

Date: 2018-03-13
Location: Zoom conference

Purpose

This document is a summary of the outcome from IT security discussion between the PhenoMeNal H2020 project (http://phenomenal-h2020.eu/) and NeIC-Tryggve2 (https://neic.no/tryggve2/) experts as well as other invited experts. The objective was to discuss the security approach taken by PhenoMeNal, to evaluate current status, identify the biggest risks, and to offer advice on future activities to strengthen the security in the project components. The meeting was attended by 26 participants, listed in Appendix 1.

Background materials

Before the meeting, background document describing the PhenoMeNal Security approach was sent out via email to all registered participants: Security approach in PhenoMeNal (pdf).

Background and introduction

PhenoMeNal (http://phenomenal-h2020.eu/) is a European H2020 project that provides the means for users to instantiate Virtual Research Environments (VRE) on top of public and private cloud providers. These VREs are based on tools being available as software containers, which are orchestrated by Kubernetes. For chaining individual tools into analysis pipelines, scientific workflow systems like Galaxy, Luigi, Jupyter, and Pachyderm are used.

Tryggve2 (https://neic.no/tryggve2/) is the Nordic collaboration for sensitive data funded by NeIC and ELIXIR nodes of participating countries. Tryggve develops and facilitates access to secure e-infrastructure for sensitive data, suitable for hosting large-scale cross-border biomedical research studies.

Discussion summary

In the meeting, Ola Spjuth, leading WP5 of the PhenoMeNal project, gave an introductory presentation about PhenoMeNal and the security approach:
https://drive.google.com/open?id=1Xi5mDww8TXcERcm9vn06le2iBv_0SGQa

Live notes from the meeting are available at:
https://docs.google.com/document/d/1xccWIraOCaBap5wwyUMEHtUjqf0Eju_avO9rX8RD6_I/edit#

Below is a summary of the outcome from the discussions, and identified Action Points (AP).

When using cloudflare does that ensure that all traffic comes through cloudflare and that all other traffic is blocked?
Using cloudflare is optional in PhenoMeNal, but if you use it then traffic is limited to cloudflare.
AP: PhenoMeNal could set up CIDR rules at the cloud provider that asks traffic to come only proxied from cloudflare.

Could security be improved with Galaxy using x509 client certificates?
PhenoMeNal could generate own client certs only trusted by their services. NeIC does this for its internal services (such as internal wiki), for users that do not have an easy way of getting a certificate.
Questions raised: Would it be simple for end users to use this, or would the end user have to be an expert? Can it be done in a web browser?

AP: Joel will put PhenoMeNal in contact with NeIC x509 CA.

Where is the data to be analysed? How is it stored, encrypted/decrypted? If encrypted, who does decryption and at which point? Unencrypted data stored anywhere at any given time? Key handling?
VREs use block storage, and data can be ingested from EMBL-EBI MetaboLights, from cloud object storage, or uploaded by user. Current PhenoMeNal policies do not encourage data sharing from the VRE. When cluster is destroyed, cloud volumes are destroyed. Any data to be kept has to be explicitly saved to separate storage beforehand.
AP: Consider delete/nullify data since cloud providers do not cryptographically delete data after volumes are destroyed.
AP: Write a guideline for data operations

Are there guidelines for when to consider data "fully anonymized"? My guess is that some "anonymized" data in reality is pseudonomized. GDPR doesn't concern the former, but the latter is still personal data.
Users are instructed to only use "fully anonymized data". This is part of terms of use, which link to definitions between "fully anonymized" and pseudonymized data. Pseudonymized data must only be used on VREs spun up on networks that are properly protected (eg within institutional firewalls).
AP: Update PhenoMeNal ELSI guidelines to make this even more clear.

When specifying the tools to use (pulled from the docker hub), to you allow for specifying to use specific versions?
When using Galaxy workflow engine, versions of containers are locked for each release of the VRE, only containers from our private registry are pulled. When using Luigi or Pachyderm, users have more control over what containers to use. This is probably how we would like to have it, with more flexibility for the power-user interfaces.
AP: Consider whitelisting containers for workflows. We anticipate that this will be quite resource-demanding.

Is there support for running the VREs without internet access? Could you download the containers and spin it up locally.
This is not supported currently, but possible to achieve.
AP: Plan for offline VREs.

Can we use Singularity instead of Docker?
Not supported by Kubernetes currently.
AP: Consider Singularity when support in Kubernetes is added.

Recommendations

Networking

1. Try to limit attack surface, for example by transparent reverse proxying.
2. Consider using VPN. This could however impact user expertise.
    a. AP: Explore commercial cloud-agnostic one-click VPN services.
    b. AP: Explore setting up using cloud provider
    c. AP: Explore option to use VPN instead of Cloudflare.
3. AP: Be more clear to users that they shouldn't use untrusted software/containers, very much to the same level of not uploading sensitive data.
4. AP: Adding support for extra levels of authentication, eg setting a session cookie.

Services inside VRE

1. AP: To largest extent, limit the exposure of the services.
2. AP: A pluggable authentication mechanism would be desirable
    a. Galaxy can use openID auth for instance, but this might be a complication for some users that might not have any accepted credentials for openID.

Container sources and build process

1. Consider signing containers if little effort is required. But given that we are contacting our registry over a secure connection, it is not of the highest priority. More important from reproducibility perspective.
2. Not all containers are built on the same hardened base image (but actually most do; e.g. ubuntu 16.04).
    a. AP: Recommendation could be to use one of 3-4 verified secure base images from dockerhub. It is possible that enforcing this would not be very disruptive.

Appendix 1: Registered participants:

- Joel Hedlund (chair), NeIC Tryggve2 Scientific manager
- Ola Spjuth, Uppsala University, PhenoMeNal work package lead
- Gianluigi Zanetti, CRS4, PhenoMeNal project
- Marco Enrico Piras, CRS4, PhenoMeNal project staff
- Steffen Neumann, IPB Halle, PhenoMeNal project, listening only
- Regina Becker, Luxembourg Centre for Systems Biomedicine (LCSB) / Uni Luxembourg ELIXIR-LUX
- Tim Ebbels, Imperial College London, PhenoMeNal project , ELSI
- Niclas Jareborg , NBIS Data manager, ELIXIR-SE, Tryggve2 staff
- Frédéric Haziza, NBIS, ELIXIR-SE Tryggve2 developer
- Ulla Rudsander, Karolinska Institutet,
- Jarno Laitinen, CSC, ELIXIR FI, Technical coordinator, cloud and data management
- Ingimar Jonsson, RHnet/University of Iceland
- Johan Viklund, NBIS, ELIXIR-SE, Tryggve2 developer
- Nanjiang Shu, NBIS, Elixir-SE, Tryggve2 developer
- Jaakko Leinonen, CSC - IT Center for Science Ltd, Sensitive Data Platform Coordinator, ELIXIR-FI, Tryggve2 developer.
- Pontus Freyhult, Uppsala Universitet, NBIS, Tryggve2 developer, UPPMAX sensitive clusters
- Minna Ahokas, CSC, Tryggve2 data coordinator
- Luca Pireddu, CRS4, PhenoMeNal project staff
- Evgeni, JohnSnowLabs, devops engineer
- Anders Larsson, Uppsala University, PhenoMeNal project staff
- Jon Ander Novella, Uppsala University, PhenoMeNal project staff
- Marco Capuccini, Uppsala University, PhenoMeNal project staff
- Matteo Carone, Uppsala University, PhenoMeNal project staff
- Samuel Lampa, Uppsala University, PhenoMeNal project staff
- Pablo Moreno, EMBL-EBI, PhenoMeNal project staff
- Ken Haug, EMBL-EBI, PhenoMeNal project staff/wp lead
- Ilja Livenson, ETAIS.ee (not present)
- Anne-Marie Bach, University of Aarhus/NeIC (not present)
- Daniel Schober, Leibniz Institute of Plant Biochemistry (IPB) (not present)
- Alessandro Sulis, CRS4, PhenoMeNal project (not present)
- Ali Syed, DTU, Heard of Computerome operations, Tryggve2 sub-project manager (not present)
- Franck Giacomoni, INRA (not present)
- Giorgio Fotia, CRS4 (not present)
- Rodrigo Barnes, Aridhia Informatics (not present)
- Mats Lindstedt, CSC (not present)