

945358-BIGPICTURE
Central Repository for Digital Pathology

WP2 - Infrastructure and database hosting

D2.03 - Report on private execution model implementation in pilot environment

Lead contributor	2 - LIU
	joel.hedlund@liu.se
Other contributors	5 - CSC 6 - UU 37 - JANSSEN 38 - BAYER

Document History

Version	Date	Description
V0.1	10 Jun 2022	First Draft
V0.2	11 Jun 2022	Comments from PMO
V1.0	14 Jun 2022	Draft for Managing Board review
V1.1	06 Ju 2022	Final Version for submission to IMI

Abstract

This report documents findings from ongoing work in Task 2.5 Pilot operation and Task 2.6 Iterative development that has been carried out since the establishment of the pilot system which itself was described in D2.02 Report on the pilot system establishment. While this work in general is described in report D2.04 Report on status of pilot services implementation, the focus of this report is on secure and private execution of untrusted and potentially confidential algorithms against potentially confidential data to produce well-defined and understandable results.

This report describes the services that have been put in place in the pilot system to enable direct access to datasets and algorithms, enabling execution of potentially untrusted and confidential code on potentially sensitive data, without disclosing information from either. These services are:

- The SD-Desktop remote desktop environment, provided by CSC in Finland, which allows Bigpicture users to use own or provided software to analyse directly accessible datasets on-platform, optionally in combination with own uploaded data.
- The Cytomine-server-as-a-service provided by NBIS in Sweden, which allows Bigpicture users to view, annotate and analyse directly accessible datasets on-platform, optionally in combination with own uploaded data.

This report also describes prospective investigative work to understand how functionalities from the grand-challenge.org platform -further described in the Results section- can be integrated into the Bigpicture infrastructure, in whole or in part, in order to support indirect access to datasets and algorithms.

Methods

The methodology that was employed in general for the work carried out in these tasks is described in report D2.04 Report on status of pilot services implementation, whereas this report focuses on secure and private execution of untrusted and potentially confidential algorithms against potentially confidential data to produce well-defined and understandable results.

Providing services for secure and private direct access to datasets and algorithms has been an early Bigpicture repository development priority, in order to enable users to develop their understanding of submitted datasets, and allow them to collaboratively inspect, analyse and enrich the datasets. This, in turn, is expected to facilitate further project work in development of data standards and algorithms.

Cytomine-server-as-a-service and SD-Desktop were selected as the first services for direct access to datasets and algorithms to be integrated into the Bigpicture private inference execution environment. Cytomine was chosen because of its central role in Bigpicture, as it is identified in the Description of the Action as the main user interface for data submission, interaction and enrichment. SD-Desktop was chosen in order to provide a secure and private Linux remote desktop environment to allow users to analyse data using any tool of their choice. These services are described in further detail in sections below.

Providing services for indirect access to datasets and algorithms has been a later priority. Here, users will not be able to download or inspect the dataset or algorithm, but all interactions are carried out through well-defined protocols with well-defined inputs and outputs which do not disclose any confidential information from the dataset or algorithm.

WP2 has investigated how to integrate the grand-challenge.org platform into the Bigpicture private inference execution environment, either whole or in parts, as further services offered to support both direct and indirect access to datasets and algorithms. The grand-challenge.org platform is described in further detail in sections below.

Cytomine-server-as-a-service

As the sensitive data archive (SDA) is deployed using helm charts on a kubernetes cluster, we needed to

develop an alternate Cytomine deployment system. The standard cytomine deployment uses docker containers in a single system, so the same containers could be reused for kubernetes deployment.

Kubernetes deployment was not possible for the default software router (job running system), as it used singularity containers. A new system was developed to allow job running in kubernetes instead and is currently deployed for testing. The Helm charts have made publicly available in a UU/NBIS GitHub repository at <https://github.com/NBISweden/cytomine-bp-helm>).

SD-Desktop

CSC has implemented a new fuse module for data access in SD-desktop, with graphical tool for mounting file systems, where actual files are visible. The functionality has been demonstrated earlier in other contexts, especially in the NeIC/ELIXIR Tryggve projects cf (<https://neic.no/tryggve/> & <https://neic.no/heilsa/>).

Usage of SD Remote Desktop requires CSC project according to CSC's processes and using a 2-factor authentication operated by CSC. User will use a mobile app for authentication.

There are limitations on who can eventually use CSC SD-Desktop and how much, but that technology may be beneficial for other environments as well. Using Cytomine inside this environment will also be piloted. CSC is still working a model where remote desktop user can execute jobs on backend HPC system. That would be essential for the indirect access environment. The process how to limit users' direct access to data has yet to be planned in autumn 2022.

grand-challenge.org

Open meetings for all interested Bigpicture partners have been organised with grand-challenge.org and the Bigpicture repository developers, in order to get a better understanding of the grand-challenge.org platform infrastructure. From here, work has been ongoing primarily in the LIU Scrum team to understand how best to integrate the grand-challenge.org services into the Bigpicture repository infrastructure. An internal presentation has been made and discussed in the Swedish Scrum teams.

Results

The Bigpicture repository is providing Cytomine-server-as-a-service and SD-Desktop as services for direct access to datasets and algorithms. These services thus allow users to work collaboratively and in private to analyse and enrich data that is under their own control. WP2 is investigating how to integrate the grand-challenge.org platform into the Bigpicture repository, in whole or in part, as further services to support direct and indirect access to datasets and algorithms. These services are described in further detail in separate sections below.

Concepts and services

This section describes overarching concepts relating to the private inference execution environment and private inference execution model, and how these relate to direct and indirect access to datasets and algorithms. It also describes what services are put in place, and what purposes they serve in the repository.

The Bigpicture repository offers a private inference execution environment, which is a secure means of providing services to users, ensuring that users can work and collaborate selectively and in private, and are only given access to data that they are entitled to.

This environment is used as a basis for development of a private inference execution model that will enable indirect access to datasets or algorithms. Here, users will not be able to directly download or inspect the dataset or algorithm, but all interactions will be carried out through well-defined protocols with well-defined inputs and outputs that do not disclose any confidential information from the dataset

or algorithm. However, Bigpicture has not yet started developing standards for how indirectly accessible datasets or algorithms should be structured, and thus, the private inference execution model currently only supports direct access to datasets and algorithms.

The Bigpicture repository is providing Cytomine-server-as-a-service and SD-Desktop as services for direct access to datasets and algorithms, thus allowing users to work collaboratively and in private to analyse and enrich data that is under their own control. These services are described in further detail in sections below.

An example of indirect access to algorithms could be offering users to try out an algorithm with own data, in a secure software-as-a-service web portal that allows the user to use the indirectly accessible algorithm, but which does not allow the user to download or inspect the algorithm.

An example of indirect access to datasets is benchmarking of algorithm correctness against known ground truth, for example against findings from a clinical trial. This is also known as "challenges", as made famous by the grand-challenge.org platform. Here, a data contributor can use the service to make a dataset indirectly accessible, so that algorithm authors can get the accuracy of their algorithms automatically assessed and quantified as compared with data and known correct answers as defined by the contributor of the indirectly accessible dataset. The algorithm author would then get quantitative accuracy performance metrics back, without confidential information from either algorithm nor indirectly accessible dataset being disclosed.

WP2 has investigated how to integrate the grand-challenge.org platform into the Bigpicture private inference execution environment, either whole or in parts, as further services offered to support both direct and indirect access to datasets and algorithms. The grand-challenge.org platform is described in further detail in sections below.

The private inference execution model is also anticipated to be useful for taking major steps toward the vision of morphological data mining, as an anticipated late Bigpicture project output. The present identified best idea for how to accomplish this is using a planned service for discovery metadata production. This service has not been designed yet, but where the private inference execution model will be used with a portfolio of generally useful AI models against all submitted datasets, thereby producing additional metadata that then can be used in discovery searches for visually distinguishing features of images in datasets, which would be very time consuming for human experts to visually characterise comprehensively, but could be feasibly automatable in a consistent and quantifiable manner using AI based image analysis tools.

Privacy parameters of the private inference execution model

Swedish data centre

A guiding principle is that a service should not have more access than absolutely necessary. This is applied from the infrastructure layer and up. All services are by default isolated and then ports and networks are opened as needed, database passwords are application specific. We also want to isolate users from each other. For interactive sessions then we have the option of deploying user specific infrastructure in cases where we are not convinced that the service is not isolated or tested enough. Non-interactive sessions will have well defined APIs with the type of data that can be extracted.

Finnish data centre

[SD Desktop](#) has two flavours depending if data transfer is enabled by the project or not. The stricter environment (Limited Edition), where the user cannot import or export files, was audited against Finnish Secondary Data Usage regulation in June 2022.

Only the members of the same CSC project can access the same virtual desktop. Principle is that only

those users who have applied access to a dataset are in the same project. A User accesses the dataset granted to their personal account, but mounted data can be considered accessible for the other project members as well.

User does not have root permissions. There is also no connection to Internet from the remote desktop. The only way to import and export data is SD Connect service. SD Connect encrypts data.

Services

Cytomine-server-as-a-service

The cytomine installer has been converted from running docker containers in a single machine to deployment as microservices in a kubernetes cluster using helm charts. This facilitates easy creation and destruction of instances, as well as allowing the system to be easily deployed on the same infrastructure as the rest of the system.

In order to adapt the system to kubernetes, a new job execution environment was needed. A preliminary system has been developed and deployed for testing. Currently the system does not have a direct connection to the sensitive data archive, but users can either upload their own data or request data from the archive to be manually delivered.

As a security precaution, Cytomine runs in a separate environment from the rest of the archive, making it more difficult to use a weakness in one system to exploit the other.

SD-Desktop

After login to SD Desktop, the user can start a pre-built computing environment (Linux OS), on-demand; available options offer the capability of doing simple statistical analysis to machine learning. User can bring applications via containers with SD Connect service.

Capacity is limited to a single virtual machine, which is enough for example viewing the images. It will be investigated how performance can be optimised. During 2022 is being designed a job submission interface to CSC's compute cluster suitable for sensitive data processing.

grand-challenge.org

Grand-challenge.org is an open source production platform that has functionality in place to support indirect access to datasets and algorithms. It also has functionality for direct access to datasets and algorithms, such as viewers and annotation tools. WP2 has investigated how to integrate this platform into the private inference execution environment, either whole or in parts, as further services offered through the platform, to support direct and indirect access to datasets and algorithms. The grand-challenge.org interface and model for inference execution is well known in the medical imaging diagnostics AI research community, and it would therefore be desirable to align the Bigpicture private inference execution model with that of grand-challenge.org.

Bigpicture has worked to integrate the grand-challenge.org platform onto the Bigpicture repository infrastructure. This work has not yet been completed, but the integration work has yielded some insights on the points regarding privacy and security, namely:

- using containers in order to isolate the components, and in the case of crashing only the containers will be affected and not the system.
- OAuth2 authentication protocol is in use to authenticate users through third party service (e.g., Google account, Github account).
- no network access and protected logs to avoid data exfiltration.
- Sandboxing of execution to protect against untrusted users and code.
- a private database and a protected database for keeping sensitive data.

Conclusion

The Bigpicture repository offers a private inference execution environment, which provides a secure means of providing services to users, ensuring that users can work and collaborate selectively and in private, and are only given access to data that they are entitled to. This environment has been used as a basis for development of a private inference execution model, which in its current state supports direct access to datasets and algorithms, enabling users to collaboratively and in private inspect, analyse and enrich submitted datasets, using the Cytomine-server-as-a-service and SD-Desktop services.

In Bigpicture, the exact policy for data protection is still open including how strictly data input and output need to be limited. There are also uncertainties for indirect access environment implementation on different environments. As a secure processing environment SD Desktop is user friendly to access. It still needs evaluation and design for indirect access. Direct access will benefit scalability to the backend system. As Bigpicture work on standards development progresses to cover also indirect access to datasets and algorithms, the Bigpicture private inference execution model will be extended to also enable indirect access to datasets or algorithms. The current bet is that this model will be very similar to that of grand-challenge.org, and for this reason, prospective work is being carried out in WP2 to investigate how to integrate grand-challenge.org functionalities into the Bigpicture repository infrastructure.