Project title: European Federation for Cancer Images

Project acronym: EUCAIM

Grant Agreement: 101100633

Call identifier: DIGITAL-2022-CLOUD-AI-02

D5.11: Interim set-up of local nodes for data federation (progress report for T5.5)

Responsible Partner(s): MAG

Author(s): Paris Laras (MAG), Gianna Tsakou (MAG), Stefanie Charalambous (MAG), Sofia Theodoridou (MAG), Ignacio Gómez-Rico (HULAFE), Carina Soler (HULAFE), Liisa Koivukoski (BC Platforms), Laure Saint-Aubert (BC Platforms), Carles Hernandez-Ferrer (BSC), Melanie Sambres (LIMICS), Mirna El Ghosh (LIMICS), Valia Kalokyri (FORTH)

Date of delivery: 28/06/2024

Version: 3

Table of Contents

List of abbreviations and acronyms

AAI = Authentication and Authorization Infrastructure

API = Application Programming Interface

BIOS (Basic Input/Output System)

CPU = Central Processing Unit

DIQCT = Data Integration Quality Check Tool

DR = Disaster recovery

DHCP = Dynamic Host Control Protocol

DRP = Disaster Recovery Plan

ECC = Error Correction Code memory

ELK stack = Elasticsearch-Logstash-Kibana Technology Stack

FAIR = Findability Accessing Interoperability Reusability

HVAC = Heating, Ventilation, and Air Conditioning

IAM = Identity and Access Management

IP = Internet Protocol

LTS = Long-Term Support

RAID = Redundant Array of Independent Risks

RFID = Radio Frequency Identification

RPO = Recovery Point Objective

RTO = Recovery Time Objective

SaaS = Software as a Service

UEFI = Unified Extensible Firmware Interface

UPS = Uninterruptible Power Supply

VPN=Virtual Private Network

VO=Virtual Organization

TLS=Transport Layer Security

WP = Work Package

1. Introduction

EUCAIM leverages the developments on other Research Infrastructures and repositories for health data, extending, integrating and customising services that are key for metadata cataloguing, federated search, access negotiation, data storage and computing-intensive processing.

These services have been adapted and customised to deal with the specificities of EUCAIM and to enable their integration. This document addresses this need by providing specifications Data Holders regarding the procurement, installation, and configuration of Local data nodes that are meant to integrate with the EUCAIM infrastructure.

1.1. Purpose & Scope

The purpose of this document is to provide detailed specifications and deployment guidelines for installing and configuring an EUCAIM data node server. This guide aims to ensure a smooth and efficient procurement and installation process by outlining each step required to prepare the environment, install the software, and configure necessary settings.

This document is intended for IT administrators, systems engineers, and technical staff responsible for setting up, configuring, and maintaining federated nodes within the EUCAIM project, either from organizations who are already part of the EUCAIM federation, or new applicants. It provides detailed instructions and guidelines to ensure seamless integration, optimal functioning, and continued availability of the federated nodes.

Additionally, it serves as a reference for organizational stakeholders and decision-makers involved in infrastructure procurement and management to be informed on the technical requirements and best practices necessary for effective participation in the EUCAIM project.

A minimum technical understanding of IT systems and modern programming and virtualization practices is required to fully comprehend the contents of the document. Also, it's important to note that the contents of this document include the interim version of the specifications for the local node setup, with further updates and changes planned to the document for the following stages of the project. Furthermore, the specifications presented here are expected to evolve with extended flexibility for various types of infrastructures as the project progresses, providing further details and specifications to data holders that want to set up a local data node.

Additionally, hardware requirements are provided to inform Data Holders of the expected processing capabilities of their local nodes, helping them decide whether to use existing resources or to procure new ones to meet the outlined requirements.

Finally, maintenance and configuration procedures are also described within the document, accompanied by descriptions of best practices, with the goal of outlining suggested behaviors and procedures for the long-term health and performance of the local data nodes.

1.2. Relationship with other Deliverables

The document is aligned with the following deliverables by including information relevant to the local data nodes on topics such as the Data Model, Data Pre-processing tools and Federated services. The original information related to these topics was collected from these deliverables and included in the scope of this document to provide the required context for the reader to comprehend the scope of activities this document describes, as well as to inform them of where further, more detailed information on each of these topics can be found.

1.2.1 D5.2: The EUCAIM CDM and hyper-ontology for data interoperability: initial version (outcome of T5.2)

1.2.2 D5.4: Data Pre-processing Tools and Services (outcome of T5.3)

1.2.3 D4.5: First Federated Core Services

1.2.4 D4.9 Central Core Infrastructure Set up

2. System Specifications

This section describes specifications and requirements for the procurement, installation, integration, and management of the EUCAIM local nodes by Data Holders.

The technical requirements for the integration in the federated infrastructure are structured based on the tiers of participation for data providers in EUCAIM:

- Tier 1: Compliance with the metadata model for describing the datasets available to EUCAIM.
- Tier 2: Direct (through adoption) or indirect (through transformation by a mediator component) compliance with the data model for searching purposes (a mediator component may be necessary to adapt the query syntax of the federated search to the local node searching API).
- Tier 3: Direct (through adoption) or indirect (through transformation by a mediator and a data materialisation component) compliance with the data model for processing purposes.

The three tier levels described above are related to the following federation concepts:

- Tier 1: The datasets hosted by the local node are registered in the public metadata catalogue. Ideally, this is done through the exposure of FAIR Data Points used by the central catalogue to harvest the dataset's metadata. However, manual registration of the datasets will be supported during a preliminary phase.

- Tier 2: The data of the federated node is searchable through its local searching service, which is queried by the EUCAIM federated search system through a Query Mediator component that transforms the query from EUCAIM's model to the local model and vice-versa for the results. In case the local model already complies with EUCAIM's one, the mediator component is still necessary for transforming the results and utilising the network communication middleware.

- Tier 3: in addition to the above functionalities of Tiers 1 and 2, the federated node has a data materialisation component that makes the data available for federated processing, according to EUCAIM's model.

2.1. Hardware Requirements

Hardware requirements are divided in different tiers, as they have been defined in deliverable D5.1 related to the scope of EUCAIM participation/integration that the Data Holder will carry out. Only two tiers are considered in the scope of this document: local nodes for Data Searching (Tier 2), and local nodes for Federated Processing (Tier 3). More information on EUCAIM's Tiers can be found in D4.3, which will be further updated in D4.4.

Local Node (Tier 2):

Data holders joining the EUCAIM federation as a Tier 2 participants are required to obtain and set-up a Local Node server which is capable of handling multiple requests and providing the EUCAIM federated infrastructure with efficient data querying functionality.

The hardware requirements for Tier 2 local nodes are described below. These requirements were selected according to the expected workload demands for data federation but may be updated in the future, following feedback that may be received by data holders once the first Tier 2 data nodes are established and become fully operational.

Hardware	Minimum	Recommended
CPU	8 Cores / 16 Threads 2.5GHz+	16 Cores / 32 Threads 3.0GHz+
RAM	32 GB ECC	64 GB ECC
Power Supply	Depends on the selected CPU and form factor of the Node.	

Table 1: Hardware requirements for Tier 2 local nodes

Local Node (Tier 3):

Data holders joining the EUCAIM federation as a Tier 3 participants are required to obtain and set-up a Local Node server which is capable of handling workloads related to data searching, as well as parallel data processing, and AI model training, which means that a minimum of GPU-accelerated computing power must be available.

The hardware requirements for Tier 3 local nodes are described below. These requirements were selected according to the increased workload demands for federated processing. They may be updated in the future, following feedback that may be received by data holders once the first Tier 3 data nodes are established and become fully operational.

Additionally, specifically for GPUs, if the format and slots of the local node allow it, the installation of multiple GPUs is possible, to allow running concurrent GPU processing jobs, massively increasing the performance throughput of the local node. In that case, it is suggested to procure the same GPUs multiple times, and to focus on obtaining maximum VRAM for the selected model.

Hardware	**Minimum**	**Recommended**
CPU	16 Cores / 32 Threads 2.5GHz+	32 Cores / 64 Threads 3.0GHz+
RAM	64 GB ECC	128 GB ECC
GPU	>150 Tensor Cores 16 GB VRAM	>300 Tensor Cores 24+GB VRAM
Power Supply	Depends on the selected CPU and form factor of the Node, as well as the added power consumption by installing one or more GPUs.	

Table 2: Hardware requirements for Tier 3 local nodes

2.2. Storage Requirements

Any Data Holder that intends to participate in the EUCAIM federation with the purpose of data sharing, and possibly federated processing, must ensure that appropriate storage drives are installed on the Local Node. These storage requirements provided below aim to:

- Allow ample storage space for shared data, including local data transformations required for each of the 3 Tiers.
- Provide additional space for the configuration of redundancy measures, such as RAID.
- Provide Tier 3 Local Nodes with additional space to temporarily store project files and metadata as required for federated processing.

Hardware	**Minimum**	**Recommended**

Operating System Drive	240GB SSD	480+GB SSD
Data Storage	2x(Dataset size) HDD or SAS Drives	4x(Dataset size) HDD or SAS Drives
RAID Configuration	RAID 5	RAID 5
Spare Drives	1x(Dataset size)	2x(Dataset size)
Power Supply	Make sure to include the amount of Storage Drives selected in the calculation of required Wattage capabilities by the Node's power supply.	

Table 3: Storage requirements

As an example, taking the above table into account, a Data Holders that intends to share a data set measured at 4 Terabytes:

- Is required to provide a 240 GB or 480+GB SSD drive for exclusive use by the local node's operating system.
- Is allowed to provide a minimum of two 4 Terabyte drives for data storage (8TB total). This is provided as a bare minimum, narrowly allowing a RAID 5 configuration between the drives as a redundancy measure. In this case, no additional space will be left for secondary data, metadata or local data transformations.
- Is recommended to provide at least four 4 Terabyte drives for data storage (16TB total). This configuration allows for additional storage space for secondary data, and further configurations, as well as space for local data transformations. Furthermore, with added space, Data Holder participating in EUCAIM's Tier 3, would be able to store, backup and manage Federated Processing metadata and artifacts independently from the operating system drive.

As such, it is suggested that data holders procure and provide redundancy for 2-4 times the storage volume required for storing the dataset locally, as this will create flexibility for future activities and reinforce the capabilities of the EUCAIM federated infrastructure.

Data Holders are also recommended to pursue the highest degree of infrastructure elasticity whenever possible when procuring and setting up this infrastructure. This recommendation for elasticity is future-facing, as Data Holders might be required to scale their infrastructure either horizontally (by adding more local nodes) or vertically (by increasing the storage capabilities of specific local nodes). These requirements might result from the future steps of EUCAIM or through a Data Holder's participation in other research and innovation projects, who consequently will require more storage space to generate new datasets.

2.3. Network Requirements

To enable the functionality of the EUCAIM local nodes, the local node must be supplied with continuous, stable access to the Internet. To allow the efficient connection of the local node to the Internet and to guarantee that the connection's bandwidth does not create a bottleneck for the operations of the node, a minimum connection of 200Mbps connection is suggested for installation of the node. Ideally, the connection are recommended to be symmetrical, and able to provide both 200Mbps of upload, and 200Mbps of download speeds. The use of research and academic networks connected to the GEANT federation[1] is suggested.

3. Hardware Installation

This section is dedicated to outlining basic instructions and best practices for hardware installation. If the reader already has a complete understanding of basic hardware management, installation and configuration for modern IT systems, it is suggested that they skip reading this following section, and proceed to Section 4, Configuration Steps.
Furthermore, a wide range of technical expertise is required depending on the topic of each section, so on a case-by-case basis it is advised to assign responsibility for specific topics to specific personnel.

3.1. Installation Guidelines

Unpacking and Inspection:

Upon receiving dedicated hardware for the local node, the data holder organization is recommended to carefully unpack each component. Open the packaging with caution, ensuring nothing is discarded. Check the contents against the packing slip and invoices to confirm all items are present. Inspect each item for any damage from transit, including dents, tears, scratches, or cracks. If any damage is found, contact the supplier immediately and document the issue with photographs of the packaging and the damaged items.

Physical Installation - Server Rack Format

When deploying rack-mounted servers, it is essential to ensure that the server racks are securely installed within the designated server room or data center. Use appropriate mounting kits specifically designed for the rack model being used and take care to properly align each server unit within the rack. Each server must be securely fastened to prevent any unintended movement or dislodging, which could potentially lead to hardware damage or operational disruptions.

Additionally, ensure that the server room or data center is equipped with adequate climate control and ventilation systems to maintain optimal operating temperatures for the servers, and that the

[1] GÉANT Network https://network.geant.org/

server machine is installed in a secure, access-controlled environment, restricted to authorized personnel only, and with emergency systems available such and fire extinguishing systems

Physical Installation - Office Computer Format

In scenarios where consumer-grade office computer hardware is used, such as high-performance rendering or gaming computers, similar guidelines for secure installation must be followed. The server computer should be positioned on a stable, level surface that can support its weight and prevent any potential tipping or movement. It is crucial to place the computer in a secure location, away from high-traffic areas or moving objects, to avoid accidental physical damage.

Power Supply Connection:

During installation, Data Holders will have to connect the server to a reliable power source, preferably supplied with an Uninterruptible Power Supply (UPS) to protect against power surges and outages, or, alternatively, a voltage stabilizer would also provide a baseline quality of electrical current to the server. During installation ensure that power cables are neatly organized and labeled to avoid any confusion or disconnections during maintenance.

Network Connectivity:

It is suggested to connect the server(s) designated for the federated node to the network using a wired Ethernet connection. This approach guarantees stable and high-speed internet access, which is essential for maintaining the functionality required for integration with the federated infrastructure of EUCAIM.

Begin by verifying that all network cables are securely plugged into the correct ports on both the server and the network switch or router. Proper insertion is crucial to prevent any potential connectivity issues that could arise from loose or improperly connected cables. Following physical connection, ensure that all necessary network configurations are meticulously completed. This includes the assignment of static Internet Protocol (IP) addresses or the configuration of Dynamic Host Configuration Protocol (DHCP) to dynamically allocate IP addresses, depending on your organization's network infrastructure and existing policies.

Initial Boot-Up:

Upon powering on the server(s) for the first time, it is crucial to verify that the boot sequence completes without any issues. Begin by observing the initial boot process, ensuring that all connected hardware components are detected and initialized correctly. During this phase, it is essential to access the BIOS/UEFI settings to confirm that the hardware, including the CPU, memory modules, storage devices, and network interfaces, are properly recognized and configured according to the expected specifications.

Navigate through the BIOS/UEFI menus to ensure that all hardware components are functioning as expected and that any necessary configurations, such as enabling virtualization support or setting the correct boot order, are appropriately adjusted. This step is fundamental to guarantee

that the server hardware is optimally configured for subsequent software installations and the effective operation of the federated node.

Hardware Documentation:

Data holders are suggested to keep detailed records of the hardware installation process, including configurations, serial numbers, and warranty information. Continue to document any changes or upgrades to the hardware setup for future reference.

3.2. Best Practices

Environmental Controls:

It is imperative to maintain an optimal temperature and humidity level in the room where the server is installed to prevent overheating and hardware damage. Ideally, the ambient temperature should be kept between 20°C to 30°C and the relative humidity should be maintained between 40% and 60% to ensure a stable environment.

Proper ventilation and air circulation around the servers are essential to avoid heat buildup, which can lead to hardware failures and reduced lifespan of the equipment. Position the servers in a manner that allows unrestricted airflow, ensuring that intake and exhaust areas are not obstructed. The use of cooling systems such as Heating, Ventilation, and Air Conditioning (HVAC) systems for computers units or dedicated server room cooling solutions is recommended to maintain a consistent temperature.

Furthermore, the server should be placed in an area away from potential sources of water damage, such as plumbing systems, to mitigate the risk of leaks or flooding. Avoid positioning the equipment directly under or near sprinkler systems unless absolutely necessary, and if so, ensure adequate protective measures are in place. Electrical infrastructure, including power outlets and wiring, should be inspected regularly to prevent any risk of electrical fires or surges.

The server should also be kept at a safe distance from any sources of dust, chemicals, or cleaning agents that could potentially damage the sensitive hardware components. Implementing a policy of restricting food and drink in the server room can further help in protecting the equipment from accidental spills.

In addition to these precautions, it is advisable to install the server in a physically secure location, ideally within a locked cabinet or dedicated server rack that provides both physical protection and organized cabling. This setup not only enhances security but also aids in maintaining an orderly environment that facilitates easier maintenance and troubleshooting.

Cable Management:

Implement good cable management practices to prevent tangled and cluttered cables, which can lead to connectivity issues, disconnects and make maintenance difficult. Use cable ties and labels to organize and identify cables.

Physical Security Measures:

The basis of Physical security starts with housing the server machine in a secure, access-controlled environment, restricted to authorized personnel only. This can be achieved through the implementation of advanced access control systems such as RFID card readers, or PIN-based entry systems.

In addition to controlling access, given the high sensitivity of health data, it is crucial to install a comprehensive surveillance system in order to maintain a record of access events. This can include high-resolution cameras strategically placed to cover all entry points and sensitive areas within the server room. Surveillance footage is to be continuously recorded and archived to provide a reliable record of all access events. These logs must be frequently reviewed and can be invaluable for investigating any suspicious activities or data security breaches.

To further enhance physical security, consider integrating environmental sensors within the server room. These sensors can detect and alert administrators to potential hazards such as fire, smoke, humidity, or temperature fluctuations that could compromise the hardware. Furthermore, sensors equipped with alarms can be added to the server's casing in order to alert administrators or security staff towards unauthorized access events.

3.3. Operating System Installation

To ensure the effective operation of a local node within the federated EUCAIM infrastructure, it is paramount to correctly install a robust and reliable operating system on the node hardware prior to deploying any project-specific software. While Linux is not the only compatible operating system for the integration of a local node, it is the most widely tested, so initial documentation will outline installation on a Linux system. Below is a comprehensive guide outlining the best practices for installing a Linux-based operating system on your federated node. It is also advised to read through the official documentation of the Linux distribution of your choice before proceeding with installation.

Selection of Operating System

The first step in setting up your federated node is selecting an appropriate Linux distribution. Make best efforts to select stable and reliable distributions such as Ubuntu, CentOS, or Debian, known for their widespread use and strong support communities. It is advisable to choose a Long-Term Support (LTS) version of the operating system to ensure prolonged support and stability. Additionally, consider the physical configuration of your machine—whether it is consumer-grade hardware or server-rack infrastructure—as some distributions might be more optimized for specific environments.

Preparation for Installation

Once you have selected the appropriate Linux distribution, download the latest version from the official website. To create a bootable installation media, you can use open-source tools like Rufus

[2](https://rufus.ie/en/). This involves preparing a USB drive or DVD with the installation image. Ensure the media is correctly formatted and the installation image is properly copied to avoid any issues during the boot process.

Booting from Installation Media

Insert the prepared bootable USB drive or DVD into the server. Restart the machine and access the BIOS/UEFI settings, which is typically done by pressing a specific key (such as F2, F12, DEL, or ESC) during startup. Configure the BIOS/UEFI settings to prioritize booting from the inserted installation media. This step is crucial for initiating the Linux installation process.

Installation Process

With the server booted from the installation media, follow the on-screen instructions to begin the installation process. During this phase, you will be prompted to partition the hard drives. It is essential to partition the drives in accordance with recommended guidelines, ensuring there is ample space allocated not only for the operating system but also for project-specific data. Proper partitioning can enhance performance and facilitate easier system maintenance in the future.

System Configuration

Upon completing the installation, you will need to set up the root and user accounts. Ensure that strong, secure passwords are used to enhance system security. More detailed information regarding security configurations can be found in Section 9, “Security Guidelines.” Configuring user accounts correctly is a critical step in maintaining the integrity and security of the federated node.

Post-Installation Procedures

After the initial setup, update the operating system to its latest version to incorporate the most recent security patches and software enhancements. For instance, on Ubuntu, this can be done by executing the commands `sudo apt update` and `sudo apt upgrade` in the terminal. Keeping the system updated is vital for safeguarding against vulnerabilities and ensuring optimal performance.

Verification

Finally, it is imperative to verify that the operating system has been installed correctly. Check that all hardware components, such as network interfaces and storage devices, are recognized and functioning properly. Additionally, ensure that the server can establish a connection to the public Internet, as this connectivity is necessary for the installation of project-specific software and future updates.

[2] Rufus (https://rufus.ie/en/).

4. Configuration Steps

This section outlines the technical steps required to configure a local node and prepare it for the next steps of software installation. Within this section, sample commands are provided to assist in the execution of each step. The commands are preceded:

- With the symbol **'#'** if they require administrator privileges to execute
- With the symbol **'$'** if a regular user can execute them

The configuration of the EUCAIM federated nodes involves setting up directories, users, and updating system packages to ensure a stable and secure environment. Following this, software required for the deployment of EUCAIM components must be installed before proceeding to the next steps of the guide.

4.1 Initial Configuration

User account configuration

While it is suggested to configure both basic and administrator users to enhance security and manageability, the specifics depend on the organization's internal policies.

To create a Basic User for regular, daily operations with limited privileges, run the following command:

```
# adduser eucaim_basic
```

To create an Administrator User, with sudo privileges for administrative tasks, run the following commands:

```
# adduser eucaim_admin
# usermod -aG sudo eucaim_admin
```

Configuring Home and Software Directories

The reader is recommended to treat this section as a suggestion. For data holders whose organization already has a process and structure defined for managing servers and their directories, feel free to ignore the following section.

To begin, local directories are suggested to be created within the user's home directory to use for storing configurations, downloaded software or other useful files related to the installation and maintenance of the Local Node.

Use the following commands in Debian or CentOS-based Linux distributions to create a new directory within the 'home' directory of the 'user1' user and assign the appropriate user rights. If your user is named something different, replace the 'user1' text with your username.

```
# mkdir -p /home/user1/software
# chmod 755 /home/user1/software
```

Following that, use the following commands in Debian or CentOS-based Linux distributions to create a new directory within the 'opt' directory of the 'user1' user and assign the appropriate user rights. If your user is named something different, replace the 'user1' text with your username.

```
# mkdir -p /opt/eucaim/
# chmod 755 /opt/eucaim/
```

Update system packages

Update the system's package manager to ensure you have the latest software versions and dependencies. You can do this by running the following commands as an administrator user in a Linux terminal:

For Debian/Ubuntu:

```
# apt-get update
```

For CentOS/RHEL:

```
# yum update
```

4.1.1. Installing Required Software

The following section describes the steps necessary to install some basic Linux tools for downloading and accessing code repositories, as well as executing web requests. Following those steps, Docker will have to be installed as it serves as a basic software tool for deploying EUCAIM's software components in a replicable and consistent manner.

Basic Packages

Install necessary tool packages such as wget, curl, git, etc. You can do this by running the following commands as an administrator user in a Linux terminal:

For Debian/Ubuntu:

```
apt-get install wget curl git
```

For CentOS/RHEL:

```
yum install wget curl git
```

Docker Installation

Follow the following steps to install Docker, a container management software on the local node. It is suggested to instead follow the official Docker installation guidelines[3] in the case that an update in the software alters the required installation steps.

Install necessary packages to allow apt to use repositories over HTTPS:

For Debian/Ubuntu:

```
# apt-get install -y apt-transport-https ca-certificates curl software-
properties-common
```

[3] Docker installation guidelines https://docs.docker.com/engine/install/

For CentOS/RHEL:

```
# yum install -y yum-utils device-mapper-persistent-data lvm2
```

Add Docker's Official GPG Key to your Linux package management software:

For Debian/Ubuntu:

```
# curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -
```

For CentOS/RHEL:

```
# rpm --import https://download.docker.com/linux/centos/gpg
```

Add the Docker official repository to your repository list:

For Debian/Ubuntu:

```
# add-apt-repository "deb [arch=amd64]
https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
```

For CentOS/RHEL:

```
# yum-config-manager --add-repo
https://download.docker.com/linux/centos/docker-ce.repo
```

Install Docker:

For Debian/Ubuntu:

```
# apt-get update -y
# apt-get install -y docker-ce docker-ce-cli containerd.io
```

For CentOS/RHEL:

```
# yum install -y docker-ce docker-ce-cli containerd.io
```

Start and Enable Docker service

For Debian/Ubuntu AND CentOS/RHEL:

```
# systemctl start docker
# systemctl enable docker
```

Validate Docker installation

Attention: Before executing the following command, be sure to log-out and back into the system, in order to update your user's permissions to the Docker user group. In the case that the permissions are not updated, the command might not run, and you will have to execute the following command as an administrator user.

For Debian/Ubuntu AND CentOS/RHEL:

```
# docker run hello-world
```

4.1.2. Network Configurations

To properly configure the network access of the federated node, firstly configurations must be made to the network's firewall. To achieve this, the organization's technical staff must configure network firewalls to allow only necessary inbound and outbound traffic.

Notify and Collaborate:

If using additional security protocols like VPNs or reverse proxy networks, notify the Project's Technical Support team to ensure compatibility and security compliance, as these cases require specified configurations, and thorough testing before finalizing the installation.

Firewall Settings:

By default, firewalls installed on Linux machines will prevent the machine's ports from communicating with outside networks. The EUCAIM software stack requires a set of specific ports to be allowed in order for the software components to operate and communicate with the EUCAIM infrastructure. Adjust firewall settings to allow specific ports required by the federated node:

```
# ufw allow 22/tcp
# ufw allow 80/tcp
# ufw allow 443/tcp
# ufw enable
```

Furthermore, it is advisable to conduct thorough testing to confirm that the network connection is active and stable. This can involve pinging external servers, performing bandwidth tests, and checking for any packet loss. Execute the following command to verify that the federated node can reach external networks and the Internet.

```
$ ping -c 4 google.com
```

Any identified issues need to be promptly addressed to prevent disruptions. In addition, it is suggested network redundancy mechanisms be configured, such as link aggregation or failover setups, to enhance network reliability and reduce downtime risks.

5. Federated Node Integration

5.1. Connecting to EUCAIM System

To integrate the local node with the EUCAIM System, the following software must be installed and run as a service within the local node:

Beam-proxy[4]
(link in footnote)

Handling communication with Beam Broker deployed in the EUCAIM Central infrastructure, taking care of authentication, encryption, and signatures. The Beam-Proxy is responsible for connecting the application service to the broker, thereby allowing the application to exchange data and requests with other infrastructures.

Focus Query Dispatcher[5]
(link in footnote)

Receiving Beam tasks using Beam Proxy, translating queries depending on the types of endpoints, running them, and returning the results to Beam. This service is used to interface and communicate with any and all services that are being executed locally, acting as a sort of orchestrator component, taking the responsibility of communicating and returning results to the EUCAIM infrastructure.

Deploying EUCAIM integration software:

The following section contains the commands necessary to install the software. Extended descriptions of the deployment of this software are available in deliverable D4.9 “Central Core Infrastructure Set up”, which is also suggested to be read as it provides additional context for the federated search functionality and provides a full scope for understanding the EUCAIM infrastructure.

To deploy these components using docker:

1. Ensure Docker is installed on your system.
   Refer to section 4.1.1 for Docker installation instructions.
2. Create a Docker Network
   Create a dedicated Docker network for the services to communicate using the following command:

   ```
   $ docker network create eucaim-network
   ```

3. Deploy Beam-Proxy:

   Create a Dockerfile for Beam-Proxy or use the provided one in the repository. An accurate Dockerfile would contain the following:

   ```
   FROM openjdk:11-jre-slim
   COPY . /app
   WORKDIR /app
   RUN ./gradlew build
   ```

[4] https://github.com/samply/beam

[5] https://github.com/samply/focus

```
ENTRYPOINT ["java", "-jar", "build/libs/beam-proxy.jar"]
```

4. Build the Docker image
   Note: substitute "`path/to/Dockerfile`" to the path where `Dockerfile` is stored.

   ```
   $ docker build -t beam-proxy:latest -f path/to/Dockerfile .
   ```

5. Run the Beam-Proxy container.

   ```
   $ docker run -d --name beam-proxy --network eucaim-network beam-proxy:latest
   ```

6. Deploy Focus Query Dispatcher:

   Create a Dockerfile for Focus or use the provided one in the repository.

   Dockerfile contents:

   ```
   FROM openjdk:11-jre-slim
   COPY . /app
   WORKDIR /app
   RUN ./gradlew build
   ENTRYPOINT ["java", "-jar", "build/libs/focus.jar"]
   ```

7. Build the Docker image.
   Note: substitute "`path/to/Dockerfile`" to the path where `Dockerfile` is stored.

   ```
   $ docker build -t focus:latest -f path/to/Dockerfile .
   ```

8. Run the Focus container.

   ```
   $ docker run -d --name focus --network eucaim-network focus:latest
   ```

9. Verify Deployment

   Check the status of the running containers to ensure they are up and running.

   ```
   $ docker ps
   ```

10. Configure Environment Variables:

    Depending on your infrastructure, set the necessary environment variables for Beam-Proxy and Focus to function correctly. This can be done using a Docker `--env` flag or a `.env` file. The latest configuration for environmental variables of local nodes can be found here:

    https://github.com/EUCAIM/k8s-deployments/

11. Monitor Logs:

Monitor the logs for both services to verify they are functioning correctly and connected to the EUCAIM Central infrastructure.

```
$ docker logs beam-proxy
$ docker logs focus
```

By following the above steps, your organization will be able to successfully deploy the necessary components to integrate your local node with the EUCAIM System using Docker.

5.2. Registering to Services

Following the above instructions and having deployed the necessary integration components, its services must be registered in the EUCAIM IAM systems, specifically the AAI system that EUCAIM is built around, for the Node to be fully integrated.

Broadly, authentication is performed through the membership to the EUCAIM VO Group[6] for each individual user type. The process of creating an account and requesting membership to the EUCAIM VO group is described in the dashboard site[7] and involves the manual verification of valid credentials, as well as the registration of deployed services as official EUCAIM services.

6. Software Installation for Federated Searching

The components necessary to implement this functionality are:

- Provided by the EUCAIM project, the installation of which has already been documented as part of basic integration (see Section 5 for the installation of the Beam-Proxy and Focus Query Dispatcher).
- The Mediator service, which is developed and implemented by each Data Holder and allows the integration components (Focus & Beam-Proxy) to communicate with the locally available interface for exchanging information with the data-storage systems.
  An implementation of the mediator component for connecting CHAIMELEON data holder has been integrated in CHAIMELEON Dataset service. The implementation can be found in the GitHub repository[8]. ProCancer-I has also implemented a mediator for the datasets available from this project.

Data Holders can find the manifests and some more information about the deployment of the Federated Search service at the following link: https://github.com/EUCAIM/k8s-deployments/tree/main/federated-search/eucaim.
Additionally, you will be able to contact the EUCAIM Technical Support team through the Helpdesk (https://help.cancerimage.eu).

[6] Enrolment URL for the EUCAIM VO Group https://signup.aai.lifescience-ri.eu/fed/registrar/?vo=lifescience&group=communities_and_projects:EUCAIM
[7] User's registration process in EUCAIM https://drive.google.com/file/d/1EsFYxbzgpyYKggyeKrKKw3FkVecDby8P/view
[8] https://github.com/chaimeleon-eu/dataset-service?tab=readme-ov-file#integration-with-eucaim-federated-search

7. Software Installation for Federated Processing

In order to enable the functionality for Federated processing in the local node, the following software must be installed and run as a service within the local node:

FP Daemon
Link will be available in the next version of this deliverable, D5.8.

This daemon is responsible for brokering Federated Processing requests forwarded through the Focus query dispatcher component and communicating with data materialization services which have been deployed locally. This necessitates the existence of a data-reference service, which will be able to provide metadata on the availability and location of specific types of datasets.

For deploying this component using docker, installation guidelines will be provided in the next version of this document.

8. Software Installation for Data Preparation & Management

8.1. Overview of Local EUCAIM Data Tools

To ensure their data is ready to be shared and used within the EUCAIM framework, data holders must go through some pre-processing steps. To that end, the EUCAIM consortium provides a set of tools that can be downloaded locally and run on the data to share. These local EUCAIM data tools can be applied to either imaging data, clinical data, or both.

We distinguish two categories of local EUCAIM data tools:

- The mandatory local tools: tools that need to be used by data holders on their data to ensure some level of conformity and the eligibility of the data to become part of the EUCAIM data federation. To date, no local preprocessing tool is made mandatory to use at this stage of the project; however, because de-identification of data is mandatory, the *use* of de-identification tools can also be considered mandatory. More information about anonymization legal requirements can be found within D3.6.For imaging data, EUCAIM will provide tools to ensure the proper de-identification of the DICOM tags. Furthermore, there will be additional solutions for de-identifying further aspects of the images, such as removing burnt-in text in images, or defacing solution when the patient's face may be present. On the other hand, if data are compliant with the EUCAIM Common Data Model (CDM), tools developed during the project will be provided to automatically identify potential risks on the de-identification, both in clinical and imaging data. In the following sub-section, the current tools to de-identify and ensure the proper de-identification of the data are described; more information regarding these tools can be found in D5.4:

- **DICOM anonymizer**: The tool, developed by FORTH, runs on DICOM imaging data. It takes as input a folder with one or multiple patients/cases and it performs the de-identification of the DICOM tags according with a defined de-identification profile which specifies how to process each of them. The output of the tool is the DICOM files where the corresponding tags are removed, modified or kept. The tool has also a DICOM viewer integrated which allows the user to inspect the images and the DICOM metadata information once it has been processed.
- **Wizard**: This tool is under development during the project. The goals of it are to support the finding of patient re-identification risks and propose ways to mitigate them, to raise awareness on weak points of each process, to foster a secure-by-design anonymization planning and to facilitate compliance to EUCAIM requirements and accountability obligations. The tool will take both imaging and clinical data (as far as it is compliant with the EUCAIM Common Data Model) and will provide a report with the results of its analysis and the proposed ways to mitigate the potential risks of the dataset de-identification status.

- The highly recommended local tools: Tools that would ensure eligibility of the data for the EUCAIM data federation and their conformity to specific Tiers, but may not apply to all data types, and/or may not be relevant in all cases (e.g. if source data have already been pre-processed for other research project purposes). This category of tools refers, for now, to some tools from the Data Quality assessment and data cleaning catalogue. There are currently 4 tools available for data quality that are recommended to download and use locally:
    - **DICOM File Integrity Checker**: The tool, developed by HULAFE, runs on DICOM imaging data. It performs, among other features, a quality check in terms of correct number of files per sequence, corrupt files, precise directory hierarchy, and organizes all files following the structuring Dataset>Patient>Study>Series, as required in the EUCAIM framework. It applies the desired changes to the dataset and generates a report containing information about the selected sequences, corrupted files, missing files and merged files. It applies to imaging data from any Tier, and it performs a data check for corrupted and missing files, which is highly recommended for any tier; as it will enforce the required structuring of the DICOMs into hierarchical folders which is suggested for Tier 2, and mandatory for Tier 3, (more detail on this in Deliverable 5.4).
    - **Trace4MedicalImageCleaning:** This tool, developed by DeepTrace, aims at detecting and removing text in medical images, as they may be potentially identifying. This tool is still under development by the DeepTrace team, and applies to 2D ultrasounds and mammographies only, for any Tier.
    - **Data Integration Quality Check Tool (DIQCT)**: A tool developed by AUTH, which checks the clinical metadata quality (validity, completeness), the integrity between images and clinical metadata provided, the de-identification protocol applied, imaging analysis requirements and the existence of annotation and informs the

user on corrective actions prior to data upload. This tool applies to Tier 3 data only, and is still under development by the team of AUTH.

- **Tabular data curator**: This tool, developed by FORTH, applies to tabular data such as clinical data. It identifies duplicated fields (lexically similar and/or highly correlated features), outliers, and data inconsistencies and provides options to deal with missing values. This would only be useful for clinical data provided in tabular format.

Additional tools are likely to join this list of local EUCAIM data tools in the course of the project. The full list and description of all EUCAIM pre-processing tools is available in Deliverable 5.4 “Pre-processing tools and services”.

8.2. Requirements for Installation of Local Data Tools

Technical requirements: The only requirements that must be fulfilled are the ones related to the mandatory local tools, which at the current project phase include the technical requirements needed for the installation and execution of the EUCAIM anonymizer. In the future, additional requirements will be provided based on the wizard tool needs.

As for the **highly recommended local tools**, they may have distinct requirements. The current tools available do not have heavy requirements (see table 4).

	EUCAIM DICOM Anonymizer (referenced above under section 8.1)
1. CPU/GPU?	CPU: Monolithic algorithmic implementation. No need for GPU
2. Programming language	C++ and Java. The CTP tool that does the actual anonymization of the DICOM images is developed in Java while the wrapper / driver GUI itself is developed in C++.
3. Expected RAM usage	~8GB (actual RAM requirements depend on the size of the dataset)
4. Libraries	It uses CTP anonymizer tool with version x202 (https://github.com/johnperry/CTP/releases/tag/x202) for the anonymization process, and the C++ Qt framework (https://qt.io/) for the cross-platform GUI
5. Security measures	The tool is a desktop application with a Graphical User Interface for Microsoft Windows and Apple MacOS platforms. Thus, no dockerization option is available currently. Nevertheless, after its installation it can run under a constrained (less privileged) user account. At runtime, the tool needs to have access to the file system where the input DICOM data is stored, and it saves the anonymized images in some other (configurable) output folder.

	Tabular data curator	DICOM file integrity checker	DIQCT	Trace4MedicalImages
1. CPU/GPU?	CPU: Monolithic algorithmic implementation. No need for GPU	CPU. No need for GPU	CPU. No need for GPU	GPU is used if available, otherwise CPU is used Minimum 3 GB of free HDD space required for installation and running OS : Windows 10 (version 1709 or higher) Processors (minimum): any Intel or AMD x86-64
2. Programming language	Python 3.9	Python	R, Python	Matlab, Python
3. Expected RAM usage	~16GB (actual RAM requirements depend on the size of the dataset)	Depending on the data, the tool is designed to function across various specification levels. However, processing time may increase with lower specifications. It has been tested on both low and high specification machines.	16 GB	depending on the input data, minimum 8 GB of free RAM

4. Security measures	The tool is dockerized. It only requires temporary storage allocation for temporary files in the local non-privileged user. Also, the tool is dockerized and called from a web API the user does not have access to the source code.	The tool is licensed by the provider's institution and is currently undergoing scientific publication and patentability study, so it can be operated under the EUCAIM guidelines; however, access to the source code or the binaries is not yet permitted.	The tool does not have any additional requirements to the ones required to run the docker container.	

Table 4: specific technical requirements for the highly recommended local EUCAIM tools

Legal requirements: Some tools may not authorize access to the source code. It is the responsibility of the software provider, with the help of EUCAIM partners, to make sure that the source code is encrypted in a way that protects it while it allows the tools to be run locally by an external user. In addition, some tools may need license agreement to be signed for the tools to be used. in the documentation provided any The software provider is suggested to also make clear licensing, any rules for publication, etc. that the user of the tool is obligated to follow.

8.3. Accessing the Data Tools

8.3.1. Downloading the Software

Software to be used locally is available for download from the EUCAIM marketplace. The download procedure will be provided to Data Holders and documented in a forthcoming deliverable.

8.3.2. Running the Installers

Depending on the software, there may be a need to deploy the tool in a specific local environment. Full documentation is provided for each tool which describes how to deploy the tool, how to run the tool, and what to provide as input. Some tools will only require running command lines, some others will have a dedicated user interface. The documentation provided is part of a more

extensive documentation provided by developers' teams within WP5 task force 2 and is available in Deliverable 5.4 "D5.4 Data Pre-processing Tools and Services".

8.3.3. Verifying the Installations

Because of the variety of tools available to the Data Holder, it is not straightforward to verify the installation, as it would require having a test suite to run for each tool.
As a minimum, the user may simply verify that the downloaded files are not corrupted (and that the EUCAIM servers have not been compromised), with integrity check algorithms such as SHA-256 checksum.

8.3.4. Preparing for usage of Data Tools

The information in this deliverable, detailing the local pre-processing tools used by Data Holders to prepare data for EUCAIM, is complemented by insights from deliverables D5.4 and D5.9.

To facilitate seamless data provision, a comprehensive set of training materials is provided, prepared in collaboration with WP2. The training catalogue module 3 titled "Data and Tool Provision" includes instructions on how to provide data, including the process of data (pseudo-) anonymization and data curation. When needed, individual training sessions will be available for data holders, tool providers, and research communities. Alongside the training catalogue developed by EUCAIM, training materials from each of the tools will be provided, sharing detailed information on their configuration and usage, creating an extensive training library.

8.4. Preparing for ETL & Data Integration

8.4.1. Imaging data

All source imaging data will be in DICOM, which is used as an international standard for medical imaging, that groups the information related to the image acquisition into structured and standardized attributes. Data holders will be provided with an executable tool that will extract and store all attributes from all images of their dataset in a json file, following the structuring Dataset>Patient>Study>Series.The list of attributes extracted will be accessible for federated search in Tier 2. Further processing of this list will allow its transformation to EUCAIM CDM (conversion to Tier 3). DICOMweb RESTful services will be deployed on the site of data holders for querying and analyses. For data holders belonging to Tier 1 and willing to export their imaging data to the central platform, a DICOMweb API will be used to connect to the central environment and push the imaging data there.

8.4.2. Clinical data

A composed **ETL pipeline for clinical data** will be provided, which will include at least the following:

- One node for initial ingestion and normalization with two different modes (semi-automatic and manual) for correcting detected inconsistencies**.** This includes a web user interface for execution but mainly used for monitoring of the correction and normalization process. Inputs will include XLS, CSV, etc. A corrected, cleaned, and tabulated dataset will be generated after N iterations (if data holder runs manual mode and needs iterative revisions for a subset of data points).
- One or more dataset-specific transformation nodes to EUCAIM hyper-ontology CDM**.**
- One node for a final data access interface with the Data Access Service**.**

In any case, the pipeline will be deployed as a single tool, which - in an initial configuration phase - will be able to download from git repositories and build the specific transformation nodes**.** This will all be distributed using Docker Compose.

	EUCAIM modular ETL for clinical data
1. CPU/GPU?	CPU. No need for GPU
2. Programming language	Python, Bash, Make (we may rework this dependency)
3. Expected RAM usage	Recommended: 16 GB, it depends on the data however.
4. Security measures	The tool requires privileges to run Docker Compose.

In the next iteration it will be considered to split this component into two tools: one for preprocessing & validation, and one for transforming from tabular data to the EUCAIM CDM.

9. Maintenance and Monitoring

9.1. Regular Maintenance Tasks

To maintain the health and functionality of the local node, data holder's staff must perform regular maintenance tasks to ensure the continuous efficient and secure operation of the local node. Data Holders should undertake the following tasks regularly, and as part of their routine maintenance:

- Monitor disk usage and perform necessary cleaning operations to remove unnecessary files, ensuring ample storage space is available for operational needs.
- Conduct regular security audits to check for vulnerabilities. Ensure firewalls, VPNs, and other security measures are correctly configured and updated.
- Periodically inspect hardware components for signs of wear and tear. Replace any faulty components promptly to avoid downtime.

9.2. Monitoring Node Health

The monitoring task enables knowing the status of the different EUCAIM components by making requests to the associated web services at certain time intervals. In this way, it is possible to know the status of the latest checks carried out for each service, as well as other important aspects such as the remaining time until the expiration of the website's TLS certificate.

These mechanisms can also be utilized for monitoring the health of services deployed at the local node level, as long as these services are exposed to the public internet or made available for access by the Focus service.

To achieve this, a set of rules has to be defined that, when fulfilled, trigger the execution of certain actions, such as email alerts or notifications for EUCAIM technical support teams and other technicians.

For the implementation of this functionality, different services of the Elasticsearch-Logstash-Kibana technology stack (ELK stack) have been used, which are briefly described below, as they were originally defined in D4.5:

- **Heartbeat**. This service allows defining different monitors, each one associated with an EUCAIM service. In this way, the status of a web service can be verified by making requests every certain period of time.
- **Elasticsearch**. This service defines an index where documents will be generated each time the rule defined in Figure 9 is executed.
- **Logstash**. Each time a new document is inserted in the Elasticsearch index, this service will take the information from this document and will use it to notify via email the person responsible for the service that it is down.
- **Kibana**. Service responsible for graphically representing the data and metrics produced or collected by other services (in this case, those produced by Heartbeat and Elasticsearch).

9.3 Local Monitoring Tools

To ensure the health and performance of local nodes within the EUCAIM infrastructure, the technical teams of Data Holder organizations act as the first line of defense, as such, implementing local, effective monitoring tools is essential. These tools help track various metrics, identify issues with the local nodes proactively, and maintain smooth operations throughout the lifetime of the local node. Here are some general approaches for monitoring the health of the local nodes from within the infrastructure of each organization.

System Resource Monitoring

- CPU and Memory Usage:
  Monitoring the CPU and memory usage of the local node is critical to ensure that the system is not overburdened. Tools like **top**, **htop**, and **vmstat** provide real-time insights into CPU and memory consumption, allowing administrators to identify and address performance bottlenecks promptly.

- Disk Usage:
  Keeping an eye on disk usage is important to prevent storage-related issues. Command-line tools like **df** and **du** offer details on disk space utilization, helping to manage storage resources effectively. Implementing alerts for disk space thresholds can prevent potential data loss or system crashes due to full disks.

- Network Traffic:
  Monitoring network traffic helps in understanding the bandwidth usage and identifying potential network-related issues. Tools such as **iftop**, **nload**, and **vnStat** provide real-time and historical data on network traffic, enabling administrators to optimize network performance and troubleshoot connectivity issues.

Application Performance Monitoring

- Service Health Checks:
  Regularly checking the health of running services is crucial to ensure they are operational. Tools like **Nagios**, **Icinga**, and **Prometheus** can be configured to perform periodic health checks on essential services, alerting administrators to any service disruptions or performance degradation.
- Log Management:
  Collecting and analyzing logs from various services can provide valuable insights into the system's health and help in diagnosing issues. Tools like **ELK Stack (Elasticsearch, Logstash, and Kibana)**, **Graylog**, and **Splunk** enable centralized log management, making it easier to search, visualize, and analyze log data.

- Application Metrics:
  Monitoring application-specific metrics helps in understanding the behavior and performance of applications running on the local node. Tools like **Grafana** in combination with **Prometheus** or **InfluxDB** can be used to visualize application metrics, providing a comprehensive view of application performance over time.

Automated Alerts and Notifications

- Threshold-Based Alerts:
  Configuring threshold-based alerts for critical metrics such as CPU usage, memory usage, disk space, and network traffic can help in detecting anomalies early. Tools like **Zabbix**, **Nagios**, and **Prometheus Alertmanager** can be set up to send notifications via email, SMS, or other communication channels when specified thresholds are exceeded.

- Anomaly Detection:
  Implementing anomaly detection mechanisms helps in identifying unusual patterns that may indicate potential issues. Machine learning-based tools like **Datadog** and **Dynatrace** can analyze historical data to detect anomalies, providing proactive alerts and recommendations for resolving issues.

9.4 Troubleshooting Common Issues

- **Common issues and FAQs**

In order to best assist Data Holders in setting up their nodes, sharing their data, and to address possible issues they would face, a Technical Support team has been created among EUCAIM partners. This team is composed of IT experts and engineers from 7 partners' institutions (MEDEX, HULAFE, QUIBIM, HABIA, GUMED, MAG). The Technical Support team, in coordination with the WP2 enrolment team, is in charge of providing documentation on technical requirements to Data Holders. This documentation is currently being edited. This documentation will be accompanied by a Frequently Asked Questions (FAQs) section, listing common issues and how to tackle them. This section will grow as the project moves on, feeding itself with experience with new Data Holders.

Furthermore, since a wide range of technical expertise is required depending on the topic of each section, it is advised that personnel is assigned to specific technical topics (security, network, configuration), and that these person's contact details and roles are provided to the EUCAIM Technical Support team.

- **Assistance for EUCAIM partners:**

As part of the core services available in EUCAIM, an instance of Helpdesk has been deployed. It is a ticketing system deployed by EGI partners and built on the open-source Zammad system. Support units (or "groups") have been created in order to best address any request or issue from any user of the EUCAIM platform. A support unit dedicated to the Technical Support team has been created as one of the helpdesk support groups, to assist data holders. More details on the helpdesk infrastructure and the support units are available in Deliverable D4.5 First Federated Core Services.

To seek assistance, Data Holders will be able to access the Helpdesk from the EUCAIM dashboard, via two paths:

- Access to the Helpdesk by authenticating using their Life Science AAI account, which allows access to all core services of the EUCAIM platform, including the access to the Helpdesk interface. There, data holders can create a ticket describing their issue, and assign it to the Technical Support team.

- Access to the Helpdesk via a webform on the dashboard webpage. This will require that the data holder provides contact information in the webform, in order to receive assistance. Once the form is submitted, it will create automatically a ticket in the Helpdesk instance, where the Technical Support unit will be able to access the request and address it.

Note: if the ticket is not attributed to the relevant support unit, the ticket can easily be reassigned to the right unit by support members. As per internal guidelines on ticket management, the issue will be addressed at the latest, within 48 hours, and the data holder will receive an answer by email as well as in the helpdesk interface.

10. Best Practices for Security

To preserve the security of your organization's local node, data holders are encouraged to comply with the following best practices for security:

Regular Updates and Patch Management:
Ensure that all software, including operating systems and project-specific applications, are regularly updated with the latest security patches and updates.
Utilize automated update tools to streamline the patch management process if necessary and strive to minimize vulnerabilities.

User Account Management:
Create individual user accounts for all authorized personnel accessing the local node. Avoid using shared accounts to maintain accountability and traceability.
Implement strong password policies, requiring complex passwords that are regularly changed. Use multi-factor authentication (MFA) to add an additional layer of security to user accounts.

SSH Access Configuration:
Restrict SSH access to the federated node to only those with a legitimate need for remote access. Always use SSH key-based authentication instead of password-based authentication for enhanced security. Regularly review and update the list of authorized SSH keys, removing access for individuals who no longer require it.

Network Security:
Organizations must configure firewalls to only allow only necessary inbound and outbound traffic. To achieve this, any unused ports are recommended be closed to reduce the attack surface. Furthermore, your organization should implement network monitoring and intrusion detection systems, if possible, in order to identify and respond to threats in real-time.

VPN Usage:
Data Holders are advised to make use of VPNs for remote access to ensure secure and encrypted connections to 3rd party services or external infrastructures.

Audit and Monitoring:
It's critical for organizations contributing to the EUCAIM federated infrastructure to conduct regular security audits to identify and address vulnerabilities and compliance issues. Technical teams of Data Holders are suggested to implement continuous monitoring solutions to track system performance, detect anomalies, and respond promptly to security incidents as they occur. Finally, it's critical to maintain detailed logs of all access and configuration changes for auditing and forensic purposes in the case of an incident.

Incident Response Plan:
Finally, in the case of a security incident, organizations must have an incident response plan in place. To achieve this, data holders are called to develop and document an internal incident response plan outlining the steps to be taken in the event of a security breach or other incidents. Following this, staff have to be trained based on the incident response procedures to ensure a swift and effective response to security threats if they occur.

11. Backup and Recovery

The planning, design and use of a robust data backup strategy is crucial in the context of the EUCAIM distributed infrastructure, where the environment is highly variable and there may be multiple sources of error (network availability, hardware failures, file corruption, software bugs...). To be part of the EUCAIM Federation IT infrastructure, every local node needs well-defined backup policies and strategies for disaster recovery. Thus, EUCAIM aims to ensure data integrity, availability and quick recovery from any unforeseen events, maintaining continuous operation and trust in the data-sharing infrastructure between multiple nodes with the least possible downtime.

One of the most important features when backing up data and deciding its periodicity is to consider the types of data to be backed up. All original datasets must be backed up on local nodes and be available on the EUCAIM infrastructure, but it is not so straightforward for the results obtained after processing. Results that have been generated by EUCAIM validated and versioned tools (such as pre-processed or filtered datasets, or intermediate results such as masks, among others) do not need to be permanently stored, as these results can be re-generated from the original dataset at any time. However, if the datasets have been enriched by manual revisions, such as

segmentations by radiologists or annotations validated under a research project, a backup copy of these results as well as the associated metadata is recommended to be kept.

Within these policies, the characteristics of the available hardware need to be be considered, since the amount of available space and the time it can be occupied by new copies, as well as how up to date the data is, will be very conditioning.

11.1. Creating Backups

Backups are the main requirement to safeguard data against loss, corruption, or accidental deletion. A backup is a copy of data that can be used to restore the original in case of data loss. Each local node is responsible for applying their backup policies to ensure that they have in a different environment a copy of the data they have shared to EUCAIM. This can range from backups between different disks on the same node, or ideally copies to a different server in a different location.

On the one hand, backups on different disks enhance data redundancy and speed, are cost-effective, and simple to manage (see 2.2 for more information about storage requirements). On the other hand, remote server backups protect against local disasters, enhance security, and offer continuity and scalability. In this way, nodes can protect against data loss due to hardware failure, data corruption, or other localized issues.

In order to define these backup policies, it is important to identify what is going to be saved and how often, as well as to identify the save windows. A save window is the amount of time that the system cannot be available to users while the backup operations are being performed[9]. The creation of backups is best to be executed when the system is at a known point and data is not changing. It can be distinguished between:

- Long Save Window: for environments with 8-12 hours of inactivity, a simple backup strategy involves saving all data during off-peak hours.
- Medium Save Window: with 4-6 hours of available downtime, a more complex strategy may be needed, such as incremental backups that save only changes since the last backup.
- Short Save Window: for systems with minimal downtime, a sophisticated approach using journaled objects is necessary, where only changes are saved to minimize disruption.

Taking into account all this information and in order to provide guidance to local nodes, some general recommendations can be suggested. For instance, the periodicity of backups can be as follows:

- Closed projects: First day of the month every 3 months
- Ongoing projects: every week on Friday afternoons

[9] https://www.ibm.com/docs/en/ssw_ibm_i_75/pdf/rzaj1pdf.pdf

Additionally, considering the varied needs of linux environments, the following open-source tools can be utilized for backup management:

- Rsync[10]: it is a versatile file synchronization tool that transfers and synchronizes files or directories between a local machine, a remote server, or any combination of these. It minimizes data transfer by sending only the differences between the source and destination, making it a reliable choice for routine backups with minimal bandwidth usage.
- Borg[11]: it is a deduplicating backup program designed to create secure, space-efficient backups of large and dynamic datasets. It reduces storage space by identifying and deleting duplicate data across backups. Additionally, Borg offers features such as compression and encryption, ensuring the integrity and security of backed-up data with minimal disruption, particularly in environments with stringent downtime constraints.

By customizing these backup frequencies and leveraging suitable tools like the ones explained, local nodes can effectively safeguard their data, enhancing the resilience and continuity of their systems.

11.2. Restoring from Backups

Restoring involves retrieving data from backups to return the system to its previous state. It is generally recommended that the restoration process be efficient and reliable to ensure minimal disruption. The main steps that have to be taken for each local node are:

- Identify backup sources: determine the most recent and relevant backup, taking into account the defined frequency of its creation.
- Initiate the restore process: follow documented procedures to restore data from the backup location to the primary system. The choice of backup tools, such as rsync or Borg, can streamline this process, ensuring timely restoration of data.
- Verify data integrity: ensure the restored data is accurate and complete, matching the state before the loss occurred. Aforementioned tools have the ability to accurately synchronize and deduplicate data, playing a crucial role in verifying data integrity during the restoration process.
- Resume operations: once the data integrity is verified, normal operations can be resumed.

11.3. Disaster Recovery Planning

Disaster recovery (DR) involves a set of policies and procedures to enable the recovery or continuation of local nodes infrastructure and systems following a natural or human-induced disaster, within an acceptable timeframe and minimizing data loss. A well-designed Disaster

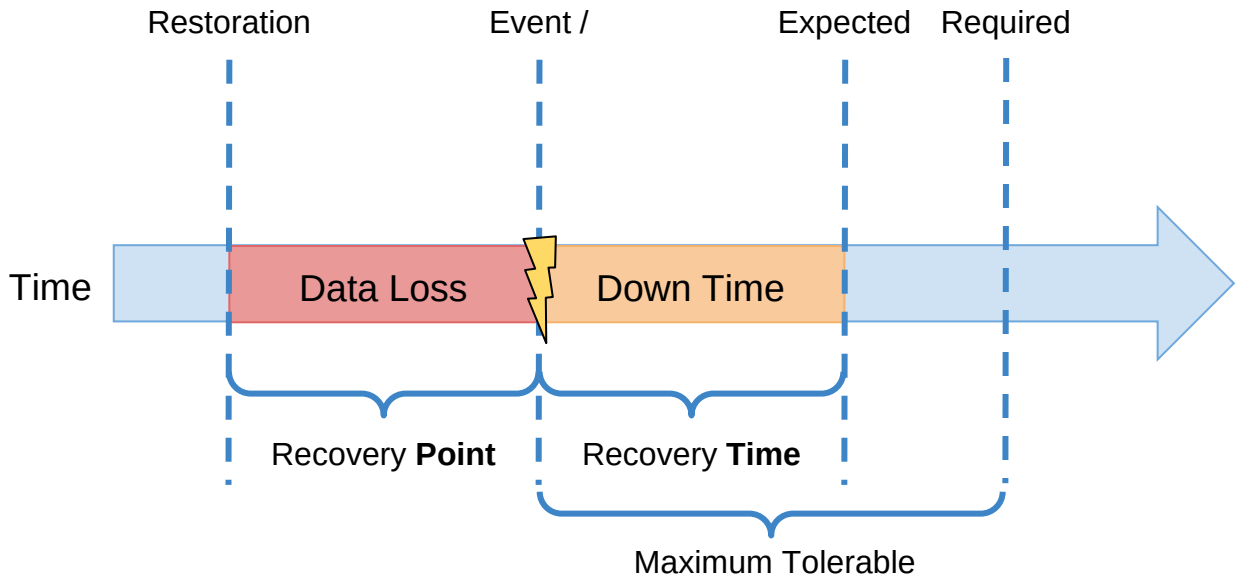
[10] https://ss64.com/bash/rsync.html

[11] https://borgbackup.readthedocs.io/en/stable/

Recovery Plan (DRP) incorporates several key elements to ensure resilience and continuity. Some of the key goals of this plan are to:

- Minimize the interruptions to the normal operations
- Limit the extent of disruption and damage
- Establish alternative means of operation in advance
- Train personnel with emergency procedures
- Provide for smooth and rapid restoration of service

An important aspect of DR planning is defining the Recovery Time Objective (RTO) and Recovery Point Objective (RPO). The RPO represents the maximum allowable amount of data loss measured in time, dictating how often backups are recommended occur to prevent unacceptable data loss. On the other hand, the RTO specifies the maximum allowable time to restore normal operations after a disaster, setting a clear target for the speed of recovery efforts.



An accurate and up-to-date hardware and software inventory is essential for effective disaster recovery. The objective of this inventory is to categorize assets into critical, important, and unimportant. Critical assets are those essential for business operations, important assets are daily used but are not critical, and unimportant assets are infrequently used. Prioritizing these categories ensures that the most crucial elements are restored first, maintaining operational integrity.

Identifying personnel roles is another vital component of a DRP. Specific responsibilities are recommended to be assigned for backup maintenance, disaster declaration, vendor contacts, and crisis management. Clear role definitions ensure that each aspect of the recovery process is managed effectively and efficiently, minimizing confusion and delays during a disaster.

The DRP must also include a list of disaster recovery sites, classified into hot sites, warm sites, and cold sites. Hot sites are fully functional data centers with up-to-date data, ready to take over operations immediately. Warm sites are partially equipped for critical systems but may not have the most recent data. Cold sites are basic facilities for storing backups without the ability to run operational systems immediately. The choice of site depends on the required level of data availability and recovery speed.

Remote storage of physical documents and storage media is crucial to prevent loss during a disaster. Ensuring that copies of critical documents and media are stored remotely safeguards against localized incidents that could otherwise result in significant data loss. The human and technical resources mentioned previously are recommended be registered in versioned documents including:

- Registration of personnel: Name, position, telephone number, e-mail address…
- List of applications/datasets to be backed up: Name of the application/dataset, category (critical yes/no), responsible person/data generator, periodicity of backups over this resource…
- Inventory of the hardware equipment available and involved in data storage/processing: manufacturer, model, serial number, guarantee/assurance…
- Log sources

Furthermore, considering the evolving landscape of technology and its impact on disaster recovery, it's worth noting that some Software as a Service (SaaS) solutions can offer a streamlined approach to data management and recovery processes. SaaS platforms, with their cloud-based infrastructure and automatic backup capabilities, can provide local nodes with scalable and reliable solutions for disaster recovery. By leveraging SaaS offerings, local nodes can ensure data availability and continuity while minimizing the complexity and costs associated with traditional disaster recovery methods.

Regardless of the approach followed, the most important point is to have all the DRP processes and resources well documented in order to guide the staff through the immediate actions needed to respond to a disaster. These procedures are best documented when step-by-step actions for system failover and recovery verification are listed, ensuring that recovery efforts are executed smoothly and effectively.

12. Conclusion

Maintaining a federated node within the EUCAIM infrastructure requires careful attention and adherence to best practices. This document attempts to guide the user from hardware procurement and inspection to performance tuning and high availability configurations, with each step ensuring seamless integration and functionality. Having read and followed the steps of this guide, the reader is expected to have successfully achieved the following:

- Planned and procured a local node for their organization according to the EUCAIM specifications and in accordance with their Data Holder Tier (2, or 3)
- Set up and configured their local node's hardware, and to have installed it in a secure and properly maintained environment.
- Installed the local node's operating system, configured basic user directories, and to have installed basic tools and software dependencies for integration
- Received specifications related to the EUCAIM data tools and access to the ones which are available
- Installed EUCAM components necessary for integration
- Read and take into advice the guidelines and best practices related to Maintenance, Monitoring, Security, Redundancy and Disaster recovery.

By following the guidelines in this document, Data Holders can plan ahead for the installation of their Local EUCAIM node, learn how to effectively contribute to the EUCAIM project, and enhance data exchange across the EUCAIM federation.