Deliverable D3.1 Cloud Infrastructure Resources

Work Package	WP 3 - GenDAI Safe & Cloud Computing Platform - Data Management & Long Term Archiving
Task(s)	T3.1 Cloud and Infrastructure Support
Author(s)	Massimo Mecella, Francesca De Luzi, Flavia Monti, Andrea Marrella, Francesco Leotta, Thomas Krause, Philippe Tamla, Andrea Leoni, Andrea Molinari, Paolo Buono, Haithem Afli, Mike Leske
Creation date	01/02/2025
Last update	14/07/2025
Version	V1
Dissemination Level	SEN

*All dates in dd/mm/yyyy format

Table of Contents

Inhalt

Executive summary

The GenDAI project aims to develop a cloud-enabled, AI-driven diagnostic platform that leverages metagenomic data to deliver precision medicine solutions. Central to this vision is the design and implementation of a secure, interoperable, and standards-compliant cloud infrastructure that supports all stages of the diagnostic pipeline—from data ingestion and processing to storage, reporting, and compliance.

This deliverable provides a comprehensive architecture of the GenDAI cloud infrastructure, detailing the technical design, resource requirements, and implementation strategies aligned with multiple diagnostic use cases. It documents how cloud-native technologies—such as container orchestration (Kubernetes), scalable storage (AWS S3, Glacier), AI compute clusters (GPU/TPU), and data compliance frameworks (OAIS, GDPR, HIPAA)—are integrated to create a robust and secure environment. These outcomes directly support the implementation of the diagnostic workflows described in WP3 and lay the foundation for downstream modules in WP4 and WP5.

The infrastructure detailed in this deliverable enables secure, scalable, and traceable data handling and computation for AI model training, diagnostics, and biomarker discovery. By supporting edge-cloud data transfer, federated learning, and runtime policy enforcement, it ensures privacy-by-design and full traceability. This architecture directly contributes to Milestone MS3 (Deployment of cloud components) and enables the future implementation of RIO2 (AI-assisted diagnostics), RIO3 (Automated data archival), and RIO5 (Compliance monitoring).

The proposed infrastructure enhances the overall impact of GenDAI by ensuring reproducibility, scalability, and compliance across institutions and use cases. It allows for interoperable cloud services that are immediately applicable to medical diagnostics and future scientific applications—whether in large-scale research collaborations or market-ready diagnostic platforms.

Future exploitation includes: (1) extending the infrastructure for real-time AI model feedback and continuous learning, (2) integrating additional omics data types, and (3) commercializing scalable, regulatory-ready diagnostic cloud modules. The technical solutions described here are ready to be adopted by both academic consortia and digital health innovators seeking to leverage cloud-based diagnostics with full compliance and scalability.

1 Introduction

The purpose of this document is to provide a comprehensive and detailed description of the cloud infrastructure that underpins the GenDAI project. This document outlines how the innovative diagnostics platform leverages metagenomic data and advanced artificial intelligence to deliver personalized medicine solutions. The primary goal is to establish a secure, scalable, and interoperable cloud-based environment that supports data ingestion, processing, analysis, and long-term archiving while ensuring compliance with international regulatory standards. The scope of this document encompasses all aspects of cloud infrastructure—from compute, storage, and networking to security, OAIS compliance, interoperability, and privacy controls—illustrating how these components collectively enable robust diagnostic workflows and innovative healthcare solutions.

1.1 Purpose and Scope

This document is intended to define and explain the cloud infrastructure components critical to the GenDAI project. It details the goals of developing a cutting-edge diagnostics platform that integrates metagenomic data and AI to support personalized medicine. The document specifies how cloud infrastructure services, including computing, storage, and networking, will be utilized to support high-throughput data processing and secure long-term data archiving. By delineating the technical and operational requirements, it provides a clear framework for how cloud-based solutions will facilitate efficient data management, regulatory compliance, and the seamless integration of advanced diagnostic technologies.

1.2 Document Organization

The structure of this document is organized to guide the reader from an introduction of high-level objectives to a deep dive into the technical and operational aspects of cloud infrastructure within the GenDAI project. The document begins with introductory sections that define the purpose, scope, and methodology, followed by an exploration of user requirements and challenges, including both functional and non-functional needs. Subsequent chapters detail the conceptual design and modeling of the system, and an in-depth overview of the use contexts is provided, illustrating various real-world scenarios. This organization ensures that readers can understand the progression from initial user requirements to the final system design, with clear guidance on how each element contributes to the overall cloud infrastructure strategy.

1.3 Methodology

The methodology employed in this document is grounded in a user-centered and iterative approach to defining cloud infrastructure requirements for GenDAI. Requirements were gathered through detailed stakeholder consultations, iterative feedback sessions, and extensive analysis of clinical and technical challenges. Conceptual models and architectural designs were developed by integrating industry-standard practices such as containerization, cloud orchestration, and scalable data processing methodologies. This approach ensures that the proposed infrastructure is both robust and flexible, capable of adapting to evolving technological needs and stringent regulatory requirements. Emphasis was also placed on interoperability, security, and compliance with standards such as OAIS, GDPR, and HIPAA, ensuring a future-proof design that supports the diagnostic pipeline from patient sampling to clinical reporting.

The cloud infrastructure requirements detailed in the next chapter are based on the use cases described in deliverable D1.1.

2 Use Case-Based Cloud Infrastructure Usage

2.1 UC1: Provide Patient Samples

The use case UC1 describes the sample collection and order management in the laboratory, previous to the sequencing and analysis. The use case is included to understand the complete laboratory workflow, but it is outside of the scope of the GenDAI project. As such there are no direct cloud resources associated with these systems.

It might be necessary however to connect the existing infrastructure of the laboratory such as the order management to the cloud resources described in the following section. As the existing infrastructure will not be deployed in the cloud, a AWS Site-to-Site VPN is required to connect the services.

2.2 UC2: Perform Diagnostic Tests

Our methodology for assessing cloud resource requirements for diagnostic testing relies on qualitative and quantitative evaluations specifically tailored to microbiome analysis workflows. We conducted comprehensive prestudies, including detailed interviews and observations in medical laboratories of their microbiome diagnostic processes, to capture computational, storage, networking, and security resource requirements necessary for sequencing-based diagnostics. Based on this, several components have been identified and document in deliverable D1.1. These components include a workflow management system (WMS) along with a scheduler and workflow management UI. Within the WMS several workflows need to be executed. These workflows utilize processing tasks for various bioinformatic tools as described below.

Our approach to deploying high-throughput computing resources and workflows involves leveraging cloud-based infrastructure optimized for processing extensive microbiome sequencing data. We utilize containerized services and workflow management systems, specifically designed to handle the computational complexity associated with high-throughput sequencing (HTS) data. Advanced metagenomic classification models and bioinformatics tools such as QIIME 2 are integrated into these workflows to ensure efficient and accurate diagnostic data analysis. We employ Amazon Elastic Container Service (ECS) to manage and orchestrate Docker containers, ensuring consistent execution environments and straightforward scalability. Workflow orchestration is managed through Apache Airflow. We are evaluating the use of Amazon Managed Workflows for Apache Airflow to potentially enhance scalability and simplify management overhead. The integration with other services will be managed through an API Gateway to ensure security and compliance.

Scalable cloud services are implemented through a containerized architecture that dynamically allocates computing resources, storage capacities, and networking infrastructure based on real-time processing demands. Load balancing mechanisms ensure computational resources are effectively utilized, enabling the system to handle rapid scaling during peaks of diagnostic activity without performance degradation.

Challenges encountered included managing the high volume and complexity of microbiome data. To address these, we develop solutions involving distributed storage systems, optimized bioinformatics pipelines, and automated load balancing strategies.

Compute Requirements

The orchestration and execution of workflows as part of the WMS requires compute resources provided by Amazon ECS. The exact compute requirements can vary from task to task and will be determined during execution of the project based on performance measurements. Some tasks might require memory optimized nodes to handle large data sets, while other tasks may benefit from compute-optimized nodes.

Storage Requirements

Storage capacity is provided by Amazon S3 as a distributed object storage systems, offering scalable, secure, and cost-effective storage for large datasets generated during microbiome sequencing workflows. Temporary data generated during processing stages might utilize ephemeral block storage, such as Amazon EBS, to facilitate rapid data access and processing efficiency.

Based on our prestudies we estimate an initial size of 300 MB per sample in raw FASTQ format. Modern sequencing platforms support processing of dozens to hundreds samples at the same time. We thus estimate the storage requirement per sequencing run to be in the order of several hundred gigabytes. Amazon S3 has no practical storage limitation and scales automatically, so that a more precise calculation of storage requirements is not needed at this point

Additional Resources

Amazon Managed Workflows for Apache Airflow resources are needed for evaluation and possible use in the future. To make the workflow management UI available for access a Elastic Load Balancer (ELB) will be needed. An API Gateway and appropriate network infrastructure is required to communicate with other components.

2.3 UC3: Analyse Diagnostic Test Data

This section details the cloud-based infrastructure and AI models employed for the analysis of diagnostic test data, ensuring robust processing, efficient deployment, and compliance with relevant regulatory standards. This directly builds upon the foundational AI capabilities and challenges discussed in Section 2.5, particularly concerning high-dimensional data handling, model robustness, explainability, and bias mitigation.

Cloud Resource Requirement Assessment

The analysis of diagnostic test data in GenDAI requires a robust cloud infrastructure capable of processing high-throughput sequencing data, deploying AI-driven inference models, ensuring compliance with GDPR, HIPAA, and IVDR, and providing scalability to manage variable workloads efficiently. The cloud-based approach offers elastic scalability, fault tolerance, and interoperability with external clinical databases and regulatory bodies. Automatic provisioning of resources ensures efficient handling of diagnostic workloads while maintaining data integrity and accessibility. This aligns with the computational and data requirements outlined in Section 2.5.3.

Compute Infrastructure for AI Model Execution

The cloud infrastructure leverages high-performance computing resources such as cloud-based GPUs and TPUs (e.g., NVIDIA A100, Google TPUs) to support deep learning inference. AI models are deployed using Kubernetes and Docker to ensure reproducibility and scalability across cloud and edge environments. Trained AI models, such as DNABERT-2 for genome sequence analysis (as introduced in D1.1), are stored in a Model Registry and accessed via a model inference API. The inference system automatically scales based on incoming diagnostic test loads, ensuring responsiveness and efficiency. Explainability tools such as SHAP and LIME are integrated to enhance model transparency and interpretability, directly addressing the user requirement for explainable AI discussed in Section D1.1.

Cloud-Based Data Processing and Feature Extraction

Diagnostic data undergoes automated quality control using cloud-based pipelines that remove low-quality sequences. AI-driven feature extraction methods, including the improved BiGAMi algorithm (as detailed in D1.1), identify key microbial features for diagnostic classification. Raw sequencing data is stored in AWS S3 and processed in parallel using Apache Spark on Databricks. Extracted features are then stored in a feature store such as Google Vertex AI Feature Store or BigQuery for future analysis and integration into diagnostic workflows, facilitating the handling of high-dimensional data and supporting robust model training and inference.

2.4 UC4: Create Findings Report

End-users will have access to the analysis reports through different media: web interface or documents, such as PDFs. Such reports will be available for further download in a private user area. The information exchange will be secured through encrypted data, but will be available to the end-user’s position or device. The logical architecture is compliant with PDFaaS, where the PDF is provided on demand. The cloud solution will implement UI with web technologies such as TypeScript/Dart, REST API based on Java, SpringBoot, JavaScript, and MongoDB/MariaDB technology.

Customizable templates will be created for the clinical pathologists to ease the production of the reports and set up the different patient/sample/report categories.

On the cloud, Node.js and Apache server, PDF manipulation libraries (e.g., PDFKit, iText), and mirroservices with serverless functions will be required. Microservice architecture may help in making the system scalable, including caching systems.
Refer to UC6: Use Findings Report for further description.

2.5 UC5: Manage & Archive Patient Diagnostic Data

To effectively manage and archive patient diagnostic data within the GenDAI ecosystem, implementing a cloud-based infrastructure ensures scalable, secure, and standards-compliant data handling. The assessment of cloud resource requirements followed a methodology that included workload simulations, historical usage analyses, and compliance audits.

This led to the identification of key needs: scalable object storage (e.g., AWS S3 Glacier, Azure Blob Archive), high-throughput data ingestion pipelines (e.g., Apache Kafka and AWS Kinesis), secure compute environments for metadata processing (e.g., AWS Fargate and Azure Confidential Compute), and robust networking with encrypted data transfer protocols (e.g., TLS 1.3).

Long-term data preservation is designed in accordance with the OAIS (ISO 14721) reference model. Integrating enterprise-grade archiving services with cloud-native backup solutions, ensures that data remains accessible, verifiable, and usable over time. We also adopt immutable storage options to prevent data tampering, with audit trails maintained via blockchain-backed logging mechanisms.

The system supports both structured data (e.g., patient metadata in relational databases) and unstructured data (e.g., raw sequencing files, PDFs, DICOM images) by leveraging hybrid storage strategies and metadata tagging. This dual-model architecture ensures traceability and data integrity throughout the data lifecycle, from initial ingestion to long-term archival.

Among the key challenges include the volume and heterogeneity of diagnostic data, alongside the need for traceable lineage and compliance with GDPR and HIPAA. To mitigate these, the development of automated data lifecycle management workflows integrated with access control policies and data retention logic is desirable. For example, diagnostic records are auto-tagged and routed to appropriate storage tiers based on access frequency and regulatory requirements.

Technically, the contribution integrates with the GenDAI system via secure APIs that expose metadata catalogs, storage usage dashboards, and data retrieval services. These components not only enhance the broader GenDAI data management infrastructure but also ensure that patient data remains secure, accessible, and compliant over extended periods.

2.6 UC6: Use Findings Report

In support of findings generation and delivery within the GenDAI infrastructure, the development of a cloud-based service to produce both static and interactive medical reports that are regulatory-compliant, clinically interpretable, and end-user friendly will be essential.

These reports consolidate AI-derived insights—such as microbiome profiles, diagnostic markers, and risk factors—into structured documents suitable for clinical use and patient communication.

The cloud system will support multi-format report generation, enabling the creation of printable PDFs, static digital versions, and interactive electronic reports. The interactive reports offer dynamic filtering, sorting, and multi-perspective visualizations that improve user comprehension without compromising the formal constraints required by legal and clinical standards. For instance, users can highlight only abnormal findings, compare values to population averages, or access explanatory tooltips contextualizing results, inspired by patient leaflets that communicate risks through relatable population statistics.

The design of a scalable backend infrastructure using containerized microservices for data retrieval, interpretation, and rendering supports these capabilities. Audit logs track each report generation and access event, ensuring transparency and traceability. Metadata tagging, versioning, and linkage to upstream analysis pipelines guarantee compliance with IVDR and GDPR, maintaining a consistent lineage between input data, processing models, and report outputs.

Integration into the broader GenDAI ecosystem is achieved through standardized APIs and shared cloud services. Reports are automatically populated with validated findings from analytical components elsewhere in the infrastructure and can be consumed by regulatory or clinical tools downstream. Interoperability enables real-time updates, centralized auditability, and responsive interactions, such as user-triggered explanations or contextual filtering, enhancing usability and regulatory accountability.

2.7 UC7: Enforce Regulatory Compliance

To enforce regulatory compliance within the GenDAI ecosystem, establishing a cloud-native framework that ensures end-to-end traceability, data protection, and policy enforcement across all data processing workflows is essential for the project.
Cloud resource requirements assessment is based on regulatory risk modeling, compliance gap analysis, and simulated access scenarios across clinical, research, and infrastructure domains. This informed the deployment of secure compute instances (e.g., Azure Confidential Compute, AWS Nitro Enclaves), encrypted object and relational storage (e.g., AWS KMS + S3, Azure SQL with Always Encrypted), and high-availability networking components with advanced firewall policies and TLS 1.3 encryption.

Robust enforcement of GDPR, HIPAA, and IVDR compliance is achieved through multi-layered access control mechanisms such as attribute-based access control (ABAC) integrated with identity providers (e.g., Azure AD, Keycloak), dynamic consent management modules, and end-to-end audit logging using tools like AWS CloudTrail and Azure Monitor. All access events and data interactions are time-stamped and cryptographically signed, forming immutable audit trails stored in compliance-ready data lakes.

The system implements runtime encryption of sensitive patient data, pseudonymization workflows, and contextual authorization policies tailored to each user’s jurisdiction and role.

For example, clinical users can retrieve anonymized diagnostic data only if access tokens are validated and regulatory conditions are met at runtime.

The platform supports hybrid deployment architectures and is capable of operating seamlessly across both cloud-based and on-premises (local) storage environments. Containerization with Docker and orchestration via Kubernetes enables consistent deployment and interoperability across heterogeneous infrastructures, supported by layered abstract network policies for secure communication. This allows institutions with strict data residency requirements to retain sensitive data locally while still benefiting from GenDAI's secure analytics and compliance tooling via federated access mechanisms.

To comply with the GDPR's right-to-be-forgotten (RTBF), the system includes data deletion workflows that allow data subjects to request erasure of their personal information across all GenDAI components. Upon validation of the request, the subject's Persistent Unique Identifier (PUI) is used to trigger targeted erasure routines across storage layers, relational databases, and audit logs—ensuring full propagation of the RTBF action without compromising referential integrity or system functionality.

One of the main challenges is balancing stringent regulatory enforcement with system usability and performance. To address this problem, developing a layered security model that combines real-time policy engines (e.g., Open Policy Agent) with caching strategies for access validation and integrated consent tracking services could be a possible solution. A centralized compliance dashboard allows administrators to monitor access requests, policy violations, and audit trails in real time.

The compliance framework is tightly coupled with other GenDAI components via secure APIs and federated metadata registries. Concrete examples of integration within the GenDAI infrastructure include the deployment of a centralized compliance monitoring layer that aggregates audit trails, access logs, and data handling events from distributed components across the platform. This monitoring system interfaces with various modules (e.g., federated learning nodes, diagnostic repositories, and AI model pipelines) ensuring consistent enforcement of regulatory policies regardless of where data is processed or stored.

In addition, the implementation of interoperable compliance APIs allows different GenDAI services to exchange validation signals. For example, by verifying whether datasets meet consent and anonymization requirements before being used in training or inference. Automated policy checks are executed at runtime across environments, enabling real-time detection of non-compliant behaviors and facilitating proactive remediation.

Moreover, standardized metadata schemas and version-controlled documentation ensure traceability of data lineage and model evolution, a critical requirement for regulations such as IVDR. These mechanisms allow the infrastructure to support collaborative innovation while maintaining strict adherence to data protection and transparency obligations across all participating nodes.

2.8 UC8: Discover Biomarkers

The identification of biomarkers in the GenDAI project requires a cloud infrastructure capable of supporting large-scale AI-driven analysis of metagenomic data. The system must handle computationally intensive tasks for AI model training and inference while ensuring compliance with GDPR, HIPAA, and IVDR regulations. The infrastructure must also provide scalable storage and high-throughput data processing to accommodate extensive biomarker discovery workflows. A cloud-based approach enables automated resource scaling for AI workloads, seamless integration with clinical datasets, and secure data management. This ensures that biomarker discovery can be conducted efficiently while maintaining regulatory compliance and scientific reproducibility.

The cloud infrastructure leverages GPU-accelerated computing environments (e.g., NVIDIA A100, Google TPUs) for deep learning model training. AI models are deployed and managed using containerised environments such as Docker and Kubernetes, ensuring flexibility and scalability across cloud platforms. Models such as DNABERT-2 and the BiGAMi feature selection algorithm are trained on large-scale metagenomic datasets. The training workflow includes data augmentation techniques, hyperparameter optimisation, and cross-validation to enhance predictive performance. The trained AI models are stored in a cloud-based Model Registry and deployed via APIs for downstream analysis.

The biomarker discovery pipeline integrates cloud-based batch processing and real-time analytics to process metagenomic sequences efficiently. Data is pre-processed using cloud-based tools such as Apache Spark and TensorFlow Extended (TFX), ensuring high-throughput and scalable data handling. AI models extract and rank candidate biomarkers based on statistical significance and predictive relevance. Feature extraction pipelines store processed data in structured repositories such as Google BigQuery and AWS Redshift for further validation and downstream integration.

Data security and compliance are enforced through end-to-end encryption, role-based access control, and GDPR-compliant storage solutions. Federated learning frameworks allow for collaborative AI model training across multiple institutions while maintaining patient data privacy. Long-term storage of biomarker-related data is implemented using hybrid cloud storage solutions such as AWS S3 Glacier and Google Cloud Storage Nearline. Automated data versioning and immutable storage snapshots ensure that research findings remain reproducible and verifiable over time.

The cloud infrastructure supporting biomarker discovery in the GenDAI project is designed to handle large-scale metagenomic data processing and AI-driven biomarker identification while ensuring scalability, efficiency, and compliance with GDPR, HIPAA, and IVDR. A comprehensive cloud resource assessment identified key requirements, including high-performance computing for AI model training, distributed storage for sequencing data, efficient networking for rapid data transfer, and privacy-preserving AI techniques. The infrastructure leverages auto-scaling GPU clusters deployed across multi-cloud environments (AWS, Google Cloud, and on-premise HPC clusters) to optimise computational efficiency for deep learning tasks. Distributed storage solutions, such as Google Cloud Storage Nearline and AWS S3 Glacier, facilitate secure and scalable data management, while federated learning frameworks enable collaborative AI model training across multiple institutions without exposing sensitive patient data. High-performance computing (HPC) is integrated into the biomarker discovery pipeline through Apache Spark for parallel data processing, TensorFlow Extended (TFX) for AI-driven feature extraction, and Google Vertex AI for scalable model execution. AI models are continuously trained and refined, with outputs stored in a centralised Model Registry, ensuring seamless integration with the broader GenDAI diagnostic ecosystem. To address computational and data security challenges, auto-scaling GPU clusters optimise resource allocation, edge computing minimises data transfer bottlenecks, and end-to-end encryption with role-based access control (RBAC) safeguards data privacy. These solutions ensure that biomarker discovery workflows remain highly scalable, efficient, and compliant with international regulations. Future improvements will focus on enhancing AI explainability, integrating multi-omics data, and developing real-time AI training feedback loops to further refine biomarker identification and support precision medicine advancements within the GenDAI framework.

3 Conclusions

This deliverable, D3.1, progresses the work of Task T3.1 within Work Package 3, fulfilling all its stated objectives. It provides a comprehensive description of the cloud infrastructure that underpins the GenDAI project. This document outlines how the innovative diagnostics platform leverages metagenomic data and advanced artificial intelligence to deliver personalized medicine solutions. The primary goal is to establish a secure, scalable, and interoperable cloud-based environment that supports data ingestion, processing, analysis, and long-term archiving while ensuring compliance with international regulatory standards.

The primary value of this deliverable lies in establishing a common, agreed-upon cloud vision that provides a single source of truth for all project partners. This de-risks future development by creating a clear and traceable reference for the future developments and prototypes. The outcomes of D3.1 fully meet the objectives set forth in the project's Description of Action.

The practical development and integration work will now proceed as follows:

- The GenDAI Diagnostics Workflow in WP2 will be based on this common infrastructure
- The development of biomarker discovery models in WP4 will be based on this common infrastructure as well.
- The UI and visualization models will be implemented to create the user-friendly GenDAI Interactive Reporting tools in WP5, again on the basis of this infrastructure.

In summary, this deliverable provides the essential roadmap that will steer the hands-on development phases.