COVID-19 Interreg Euregio Meuse-Rhine CoDaP

EUROPEAN UNION European Regional Development Fund

COVID-19 DATA PLATFORM DOCUMENTATION

Sharing data using Microsoft Azure

Contents

List of figures

About this document

This document has been produced by Ziekenhuis Oost Limburg (ZOL) to fulfil the requirement of the Interreg Euregion Maas-Rijn (EMR) study concerning a COVID-19 Data Platform (CoDaP). The study has three aims:

1. To study the heterogeneity differences in ICU patient populations;
2. To improve clinical decisions guidelines; and
3. To build an IT-infrastructure for secure and efficient data-exchange and analysis.

This document is intended for EMR project reviewers and individuals working on the study on behalf of participating institutions. All comments on this document should be directed to:

- Noëlla Pierlet – Clinical Data Manager at ZOL
- Prof. Dr. Dieter Mesotten – Anaesthetics & Intensive Care at ZOL.

Last updated	23 March 2021
Owners	Noëlla Pierlet, Ben Goethuys and Dieter Mesotten
Author	Joycelyn Harris

Introduction

The goal of the CoDaP study is to set-up an IT infrastructure that it is suitable for a multi-hospital data sharing environment. To achieve this, we must establish a way for participating institutions to share large amounts of data with each other in a secure manner. Microsoft Azure is a cloud platform that enables enterprises to share data securely with each other. It can host existing applications, utilise on-premises networks and streamline new application development. The purpose of this document is to explain to non-technical audiences how Azure underpins the CoDaP IT-infrastructure by providing an environment in which secure and efficient data-exchange and analysis can occur. It contains descriptions of Azure resources, data-safeguarding controls, and data processing mechanisms. Readers desiring a deeper, technical explanation of Azure should refer to the developer guides provided by Microsoft.

About Azure

Azure is a cloud-based solution from Microsoft that provides infrastructure, platform and software as a service to organisations and enterprises. It can host existing applications and streamline new application development. Moreover, it provides resources for collecting, analysing, integrating and transforming large amounts of data.

Benefits of Azure

A key driver for adopting Azure for the CoDaP study is its ability to handle vast datasets. It simplifies the flow of data between on-premise and cloud-based resources. Its analytics and intelligence capabilities include SQL data services, cognitive APIs and built-in support for data mining. These capabilities allow data to be analysed computationally so that patterns, trends, and associations are revealed. These key insights empower users to improve organisational decision-making and ways of working.

Azure has over 90 compliance certifications and was one of the first solutions to adopt ISO 27018, which is the international standard for cloud privacy. Over 20 of these compliance certifications are applicable in EU countries, such as GDPR and ENISA IAF. In addition, Azure's design is reinforced by the Microsoft Security Development Lifecycle (SDL). This is an industry-leading software development process that addresses security compliance during software development. Furthermore, the Azure Active Directory (AAD) service supplies Identity & Access Management (IAM) capabilities that ensure that only authorised users can access network resources. By extending the reach of pre-existing IAM capabilities, Azure enables enterprises to reduce costs and increase their organisational agility.

Interoperability is a key component of Azure. It seamlessly integrates with other Microsoft tools, such as Office 365 and Power BI. Additionally, there is a broad range of hybrid connections available that allow Azure resources to work with non-Azure resources. Azure supports internet protocols and open standards, such as XML and HTTP. Azure applications can be built with tools and languages that are familiar to many developers, such as ASP.NET and C++, and application can be deployed with little to no downtime to existing services. As such, the integrated development environment simplifies the production of cloud-based and on-premise applications.

Azure can scale according to demand. Scaling refers to how a system responds to an increase or decrease in workload. Azure can scale from serving 10 users to serving 10 million users, without requiring any additional code. Its built-in scaling features help developers to identify resources that must scale concurrently with others. Furthermore, Azure can scale automatically to ensure that capacity is available during peaks in the workload and automatically return to normal capacity when the peak drops.

Azure provides high levels of availability and minimal downtime. It is possible to maintain acceptable continuous performance when there are temporary failures in services, hardware or datacentres, or fluctuations in network traffic and workload. The backup and restore capabilities make it possible to perform disaster recoveries without relying on external support and restore data within hours following an outage.

Azure makes it possible for organisations to reduce IT hardware and maintenance costs because it supplies the infrastructure, platform and services as a service (IaaS, PaaS and SaaS). In contrast to the traditional ways of acquiring network resources, which require users to pay up front, Azure operates a pay-as-you-go pricing model. This means that organisations only pay for the resources and computing power that they need.

Azure resources

The Azure resources that are relevant to the CoDaP study are depicted in Figure 1. The Azure Active Directory (AAD), Firewall and Key Vault are resources that are concerned with safeguarding data. AAD is a service that assists with user authentication and controls access to other Azure resources. The Firewall is a managed, cloud-based network security service that prevents Azure resources from being accessed by unidentifiable or unauthorised entities. It can be scaled up or down to accommodate changing network traffic flows. The Key Vault provides a secure storage for keys, passwords, certificates and other 'secrets' required for accessing other Azure resources.

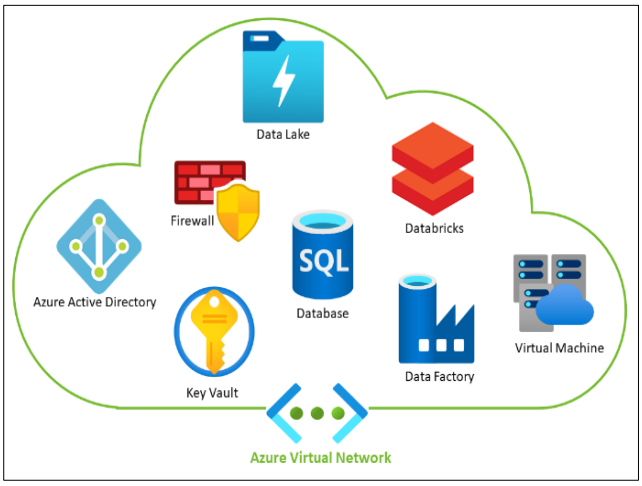


Figure 1 Azure resources relevant to the CoDaP study

Resources that are tasked with storing data are the Data Lake and the Database. A Data Lake is a resource that provides unlimited storage for structured, semi-structured or unstructured data, of any type. The Database resource stores large amounts of structured data. It is a highly available data storage layer that is based on the latest stable version of the Microsoft SQL Server database engine.

Databricks and Virtual Machines (VM) are used to analyse and integrate data. Databricks is an analytics platform based on the open-source, Apache Spark unified analytics engine. It is used for developing data intensive applications and its workspace environment supports collaboration between data engineers and data scientists. A VM is a computer file that behaves like a computer. It is used to run programs and deploy applications. A VM has its own operating system and functions separately from the other VMs. The means that multiple VMs can run simultaneously on a single piece of computer hardware. Databricks configures multiple VMs to form a cluster which work together to analyse and integrate data. Because a VM is isolated from other resources (i.e. it cannot affect their operation), it can be used for testing or backing up operating systems, accessing virus-infected data, and running applications on operating systems for which they weren't originally intended.

The central purpose of a Data Factory is to transform data. It organises integrated data into meaningful data stores that are required for informed decision-making. It provides an Extract Transform and Load (ETL) and data integration service that allows data-driven workflows to be created and scheduled. These workflows, called pipelines, assimilate data from disparate sources.

Azure resources can be grouped to form a Virtual Network known as a VNet. The VNet is a fundamental building block of a private Azure network. It insulates the resources of which it is comprised from other resources and ensures that messages passed between its resources are not intercepted. The CoDaP VNet resides within ZOL existing Azure tenant, as illustrated by Figure 2. Through filtering and routing network traffic, and end-to-end encryption, the CoDaP VNet enables secure communication between Azure resources, on-premises network resources and the Internet.

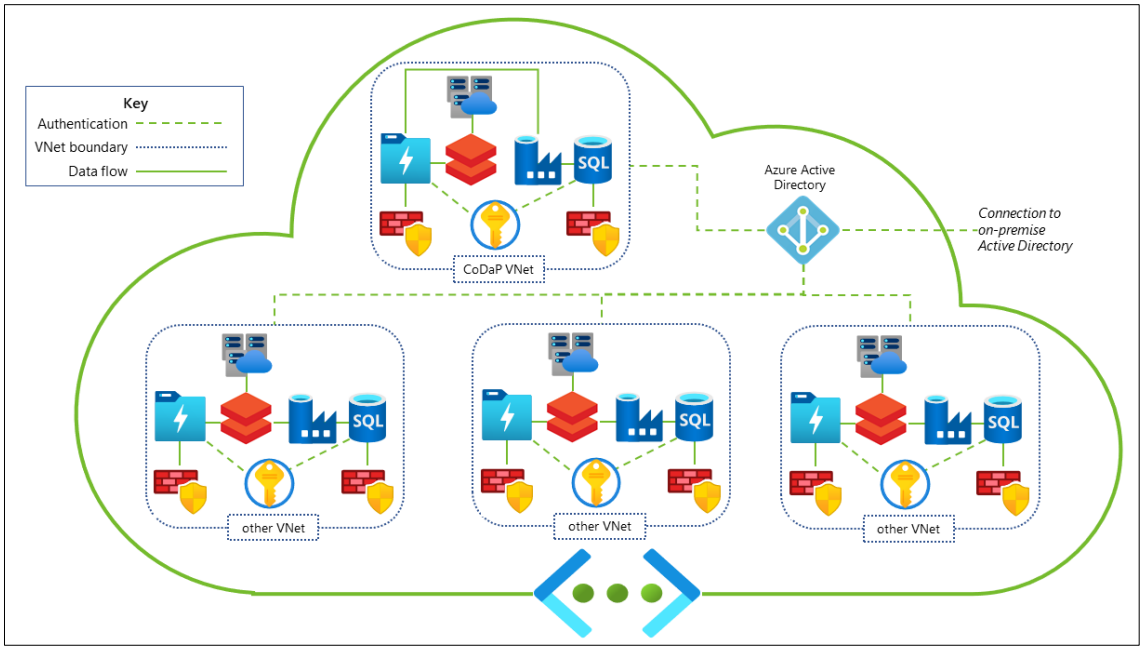


Figure 2 CoDaP VNet within ZOL's Azure tenant

The CoDaP VNet

The CoDaP VNet has been created to secure, store, integrate and transform patient data supplied by institutions participating in the CoDaP study. The tasks performed by CoDaP VNet resources are illustrated in Figure 3. They can be differentiated by whether they are concerned with safeguarding patient data or processing patient data.

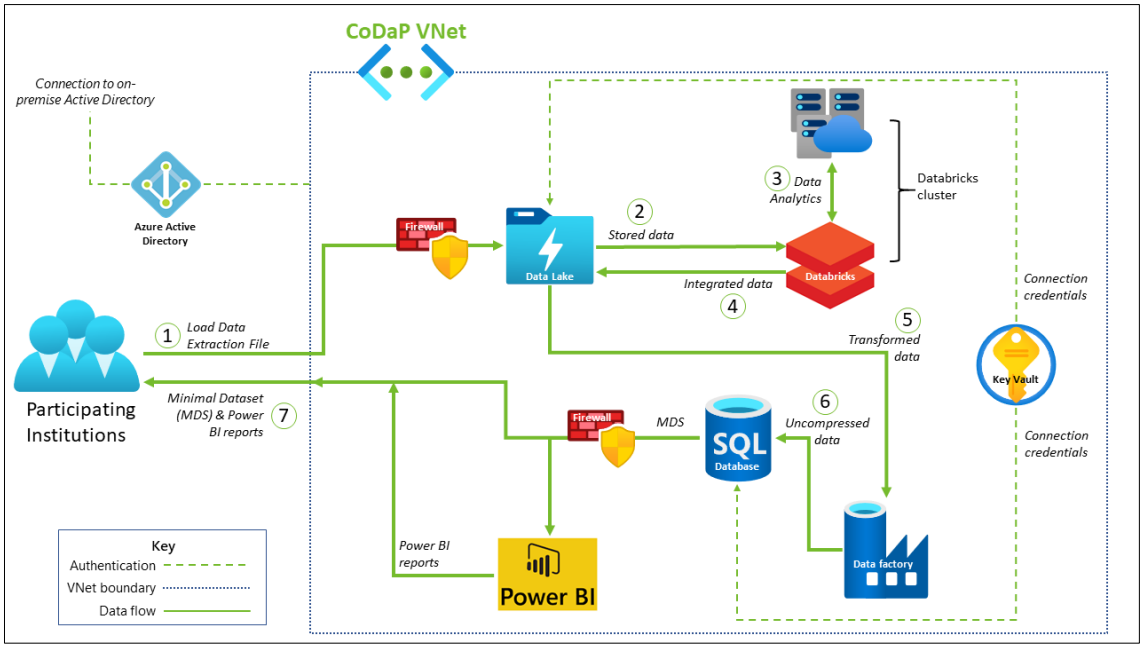


Figure 3 Overview of the CoDaP VNet

Safeguarding CoDaP data

The Azure Active Directory (AAD), Firewall and Key Vault contribute to protecting the CoDaP VNet from unwanted access attempts and malicious attacks.

Azure Active Directory

The AAD is the primary access control service. It controls the resources that a user can access and the extent. It can be synchronised with an on-premise Active Directory (AD) to authenticate on-premises users. As such, AAD is not a replacement for an existing AD service, which manages access to traditional on-premise infrastructures and applications. Rather, it extends the reach of an established AD service by enabling pre-existing resources to access to Azure resources. As previously indicated, the AAD resides outside of the boundaries of the CoDaP VNet. However, an ADD group specifically for the CoDaP study has been created, as shown by Figure 4. As the owner of the group, ZOL can make connections to the established AD directories of participating institutions. This can grant access to the CoDaP VNet to known persons or approved resources within participating institutions, whilst restricting or denying access to unapproved entities.

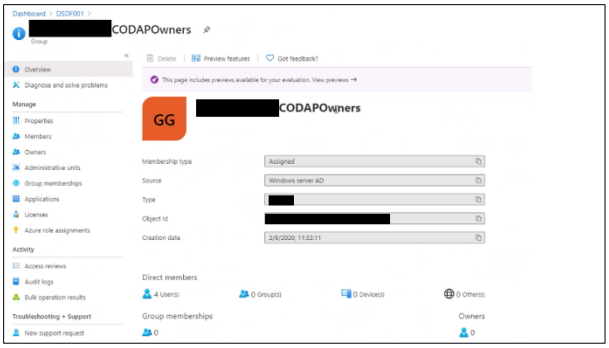


Figure 4 The CoDaP Azure AD group

Firewall

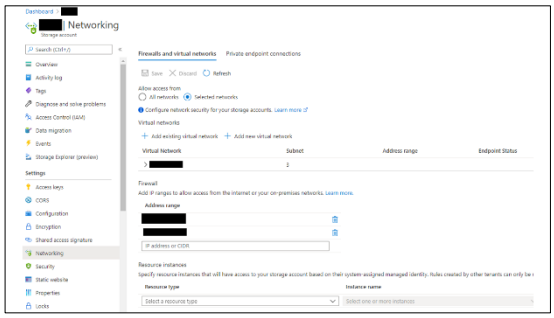


Figure 5 A CoDaP firewall

Two firewalls safeguard the CoDaP Data Lake and Database against malicious attacks and unauthorized access attempts (Figure 5). They check the IP address of a user or entity against a list of known and permitted IP addresses and only grants access if the IP address is pre-approved. If access is permitted, then the participating institution can add their patient data to Data Lake or connect to the Database.

Key vault

The Key Vault stores connection credentials required to access certain VNet resources (Figure 6). As such, it could be considered as a highly-secured password manager. If a Key Vault was not a part of the CoDaP VNet, then connection credentials would be stored directly within applications and resources as a part of their configuration. This is a risky strategy because it is relatively easy for this information to leak. By using a Key Vault, connection strings required are stored in a centralised, highly-secure location.

The key vault resides on a stand-alone hardware security module. It currently contains connection credentials that allow study participants within ZOL to access the Data Lake and Database. If an attempt is made to access the module using physical force, then it automatically deletes its contents. The connection credentials it contains are encrypted with a key that resides in the vault. Although the Key Vault can use a default encryption key provided by Microsoft, we have adopted a separate, customer-managed encryption key. This key is personal to the CoDaP study and is unknown to anyone – including Microsoft – but ourselves. Due to this, the use of a customer-managed encryption key provides an extra defence against unwanted access attempts and malicious attacks.

Processing CoDaP data

Data processing is the act of collecting data and translating it into useable information. The CoDaP study aims to brings together large amounts of data from multiple institutions and present it in such a manner that allows participants to understand and act upon it. The following resources are crucial to achieving this aim.

Data lake

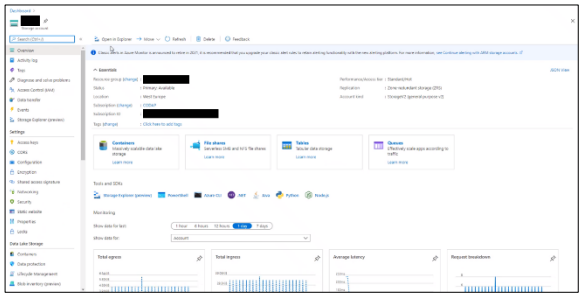
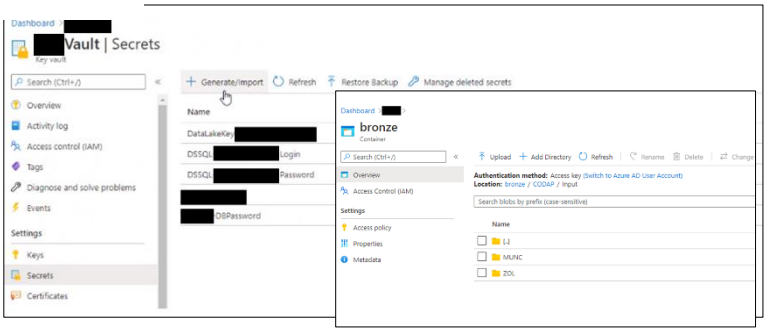


Figure 7 Data lake overview

The patient data supplied by institutions is initially stored in the Data Lake (Figure 7). The data lake is a container that can store massive quantities of data. Institutions will supply the data in the form of a Data Extraction File (DEF). The specific data that is to be provided and the structure to which it should adhere is defined in a data dictionary. More information about this can be found in the document, Data Dictionary MDS.

Unlike many other forms of container storage, the data within the data lake can be ordered in a hierarchy. In other words, folders can be used to organise the data that it contains. This is demonstrated in Figure 8, which shows that folders have been created for storing patient data originating from ZOL and MUMC+.



Figure

Figure 8 Data lake storage directory

In addition to containing DEF data, the data lake is also a repository for data that has been analysed and integrated by Databricks. This integrated data is subsequently shared with the Data Factory as transformed data.

Databricks and the VM cluster

Databricks is a control environment that configures a cluster of Virtual Machines (VMs) to analyse and integrate patient data (Figure 9). Databricks retrieves the stored DEF data from the Data Lake and instructs VMs to

manipulate it in such a manner that a single dataset is created. This dataset references data located in different data stores, such as the files and folders that contain DEFs. Databricks then loads the integrated data into the data lake.

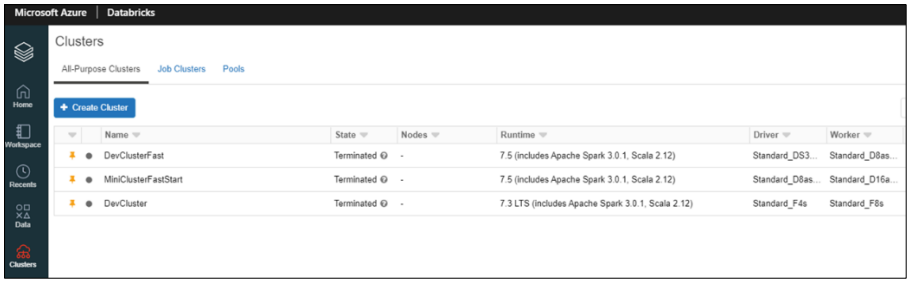


Figure 9 VM clusters in Databricks

The integration tasks that the Databricks and the VM cluster undertake are provided through Notebooks. A task a is a set of commands written in the PySpark coding language that provoke an action. Figure 10 shows some of the commands used to process patient data originating from ZOL. Because Notebooks are located in the Databricks workspace, multiple data scientists can collaborate on their development, providing that they have appropriate authorisation to do so.

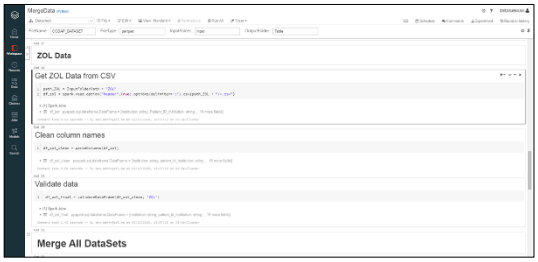


Figure 10 Commands in a Databricks notebook

Data factory

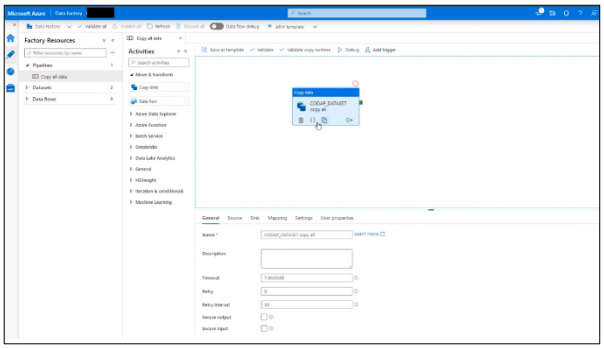


Figure 11 Data factory pipeline activities

The data factory is a production system. Its primary activity is to transform and relocate patient data from the data lake to SQL database for further processing. Its built-in ETL capabilities negate the need for the creation of custom components to transform and move data, which can be expensive and difficult to manage. Furthermore, the data factory has enterprise-grade monitoring and alerting capabilities.

CoDaP data is relocated from the data lake to the database via a Pipeline. A *pipeline* is a logical grouping of activities that together perform a task. Figure 11 shows the data factory user interface for managing the pipeline activities.

SQL database

The SQL database stores data that has been transformed and uncompressed by the data factory (Figure 12). It structures the data in such a manner that it becomes the Minimal Dataset (MDS). The MDS is the core record of the CoDaP study and consists of fields relating to demography, admission and discharge, physiology, and primary and secondary outcomes information. The MDS can be further transformed in order to gain insights that inform decision-making. MS Power BI is one software tool that can do this.

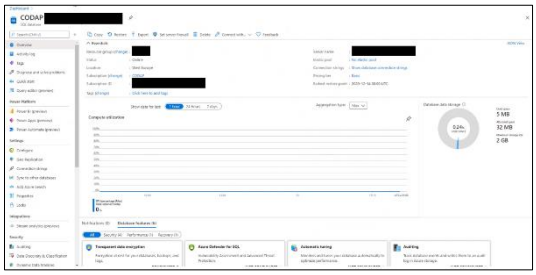


Figure 12 CoDaP SQL database overview

A description of how this is achieved is given in the documents, Reporting demographics data using MS Power BI and Reporting clinical data using MS Power BI. In the future, participating institutions may obtain MDS data by directly connecting to the SQL database.

Glossary

AAD

An initialism of *Azure Active Directory*, which authenticates users who sign in to the Power BI service.

Active Directory

A Microsoft technology used to manage computers and other devices on a network. Often initialised as *AD*.

App

An abbreviation for *Application*.

CoDaP

An acronym for *COVID Data Platform*.

Dashboard

A single page in Power BI, sometimes referred to as a *canvas*, that uses visualizations to tell a story. The visualizations on a dashboard come from reports.

Data Lake

A storage repository that can store large amount of structured, semi-structured, and unstructured data. It offers high data quantity to increase analytic performance and native integration.

Dataset

A data structure normally presented in a tabular pattern, where each column describes a particular variable and each row corresponds to a given member of the dataset. It points to or references the data stored elsewhere.

DEF

An abbreviation for *Data Extraction File*. It is the CSV file that contains datapoints (values) pertaining to patients infected with COVID-19 that have been admitted to an ICU.

End-to-End Encryption

Often initialised as *EE2E*, it ensures that messages can only be read by the sender or the receiver of the message and nobody else, including telecoms or internet service providers.

ENISA IAF

The Information Assurance Framework (IAF) is a set of assurance criteria produced by the European Network and Information Security Agency (ENISA) that organizations can review with cloud service providers to ensure that they sufficiently protect customer data.

ETL

An initialism of *Extract, Transform, Load* which is a general procedure for copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s).

GDPR

An initialisation of *General Data Protection Regulation*, which contains privacy and security rules for organizations that offer goods and services to people in the European Union (EU), or that collect and analyse data for EU residents.

Hardware

The physical parts of a computer and related devices, such as the motherboard and hard drive.

IaaS

An abbreviation for *Infrastructure as a Service*: a cloud-computing model in which a third-party provides access to computing resources such as servers, storage and networking.

IAM

An abbreviation for *Identity and Access Management*, which refers to the policies and tools used by IT departments to ensure that people and entities have the appropriate level of access to the organization's technical resources.

Institution

A hospital that is participating in this CoDaP project.

ISO 27018

A compliance standard for the protection of personally identifiable information (PII) in public clouds acting as PII processors.

MDS

An initialism of *Minimal Dataset, which* is the core record of the CoDaP study and consists of fields for that structure data related to demography, admission and discharge, physiology, and primary and secondary outcomes.

Microsoft Security Development Lifecycle

A software development process that introduces security and privacy considerations throughout all phases of the development process that enables developers to build highly secure software, address security compliance requirements, and reduce development costs.

PaaS

An abbreviation for *Platform as a Service*: a cloud computing model where a third-party provider delivers hardware and software tools needed for application development over the internet.

PySpark

A programming language used for performing exploratory data analysis at scale, building machine learning pipelines, and creating ETLs for a data platform.

Resource

A manageable item that's available through Azure. Common examples are virtual machine, database, and virtual network.

Secret

Something to which access must be tightly controlled, such as API keys, passwords, certificates, or cryptographic keys.

Tenant

The organization that owns and manages a set of Azure and Microsoft 365 services for an organization.

Vault owner

Creates a key vault and retains full access and control over it. The vault owner can also set up auditing to log who accesses secrets and keys.

VM

An initialisation of *Virtual Machine*, A computer file that behaves like an actual computer. Multiple virtual machines can run simultaneously on the same physical computer.

VNet

An abbreviation for *Azure Virtual Network*, a logical representation of your network in the cloud that allows direct private IP communication between the resources hosted in it.

ZOL

An acronym for Ziekenhuis Oost-Limburg, the healthcare provider that participating in the CoDaP study.