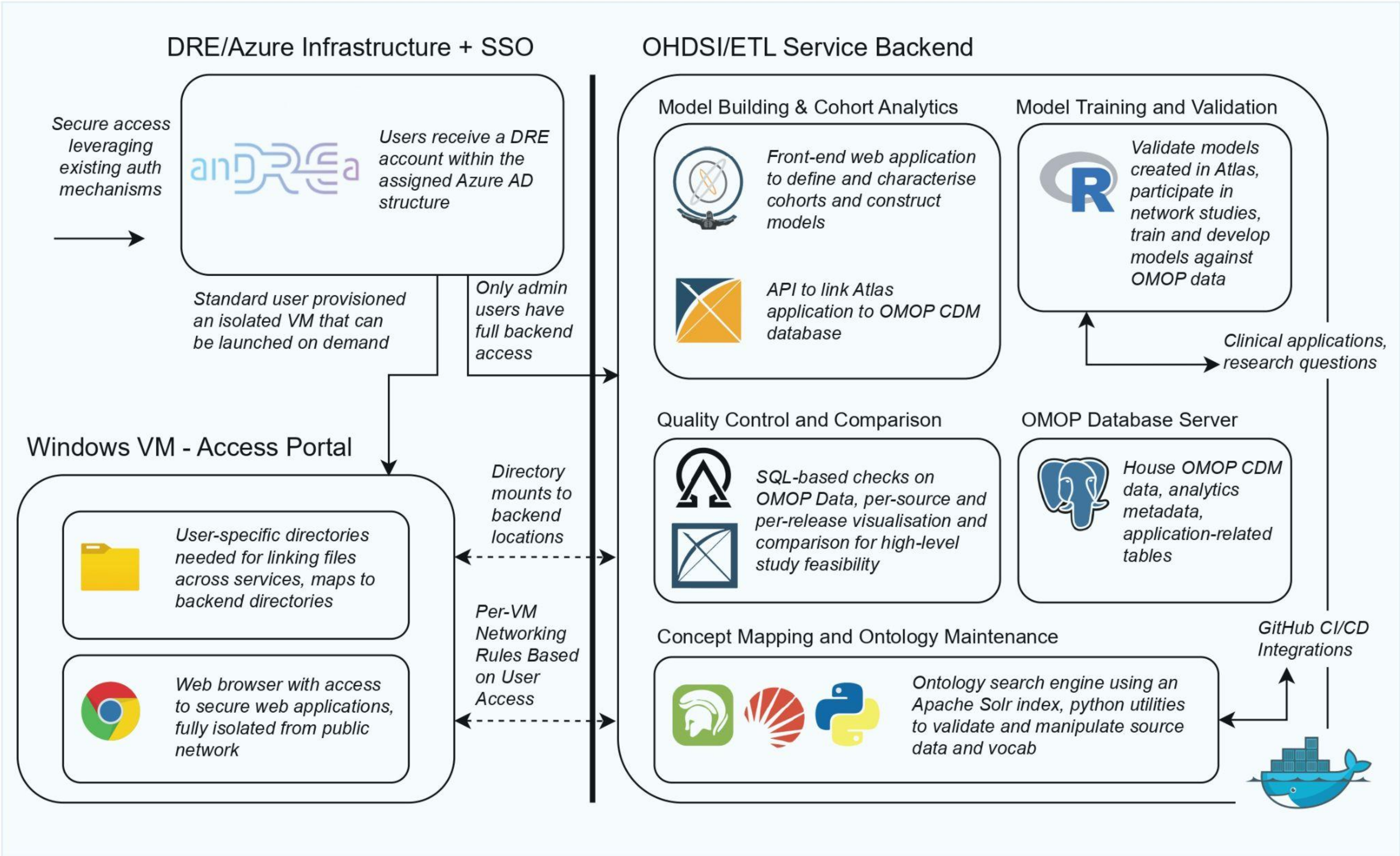
A potentially scalable, secure and cost-effective digital research environment-based service architecture to deploy and maintain ETL pipelines and OHDSI tooling for ECRAID-Base. To transform clinical data from **EU-wide, multicentre, prospective observational studies on infectious diseases and antimicrobial resistance** to OMOP-CDM

Background: European Clinical Research Alliance on Infectious Diseases (ECRAID) - Base is an EU-funded project which aims to efficiently generate rigorous evidence to improve the diagnosis, prevention, and treatment of infections, and to respond to (re)emerging infectious diseases (ID) and antimicrobial resistance (AMR) threats effectively and rapidly. At the heart of ECRAID-Base are 5 perpetual observational studies (POS). A POS is a multicentre, prospective, observational clinical study enrolling patients on a perpetual basis. Each POS creates a clinical research backbone, ready to concurrently or sequentially embed 'add-on' studies (observational, experimental, investigator-initiated, or commercial). Through EHDEN project we partnered with edenceHealth NV to transform clinical data from our POS on ventilator-associated pneumonia (VAP) in ICUs to OMOP-CDM. It was important for us to develop architecture, based on the concept of 'services-as-code' that: 1) ensures an isolated and secure workspace while preserving the privacy of patient data; 2) enables a collaborative and access-controlled virtual environment, with a degree of freedom and flexibility, for different users to work on their respective tasks concurrently; 3) provides a sustainable and cost-effective technical solution which can be quickly scaled up to ETL the other POS studies in ECRAID-Base to OMOP-CDM; 4) supports Continuous Integration/Continuous Deployment (CI/CD)



DRE/Azure Infrastructure + SSO

Secure access leveraging existing auth mechanisms

Users receive a DRE account within the assigned Azure AD structure

Standard user provisioned an isolated VM that can be launched on demand

Only admin users have full backend access

Windows VM - Access Portal

User-specific directories needed for linking files across services, maps to backend directories

Web browser with access to secure web applications, fully isolated from public network

Directory mounts to backend locations

Per-VM Networking Rules Based on User Access

OHDSI/ETL Service Backend

Model Building & Cohort Analytics

Front-end web application to define and characterise cohorts and construct models

API to link Atlas application to OMOP CDM database

Model Training and Validation

Validate models created in Atlas, participate in network studies, train and develop models against OMOP data

Clinical applications, research questions

Quality Control and Comparison

SQL-based checks on OMOP Data, per-source and per-release visualisation and comparison for high-level study feasibility

OMOP Database Server

House OMOP CDM data, analytics metadata, application-related tables

Concept Mapping and Ontology Maintenance

Ontology search engine using an Apache Solr index, python utilities to validate and manipulate source data and vocab

GitHub CI/CD Integrations

Methods:

- We **established a Digital Research Environment** (provided by Andrea Cloud - https://www.andrea-cloud.eu/azure-dre/) at University Medical Center Utrecht (UMCU) by setting up:
  - Frontend [Windows virtual machine (VM): Windows Server 2019, 2 cores, 8GB RAM]
  - Backend [Linux VM: Centos 7.5, 4 cores, 16GB RAM]
- The **ETL is orchestrated by a Python script** that executes the different SQL transformations and is packaged in a single Docker image that can be run with Docker or within a virtual environment
- The **code is version controlled using git and GitHub**. A new release of the ETL code is automatically built when a new tag is pushed to the repository
- OHDSI tooling including ARES, WebAPI, ATLAS and Athena were deployed using Docker containers
- Achilles and Data Quality Dashboard were executed together with the AresIndexer package as a single Docker process and were used for validation and iterative improvements of the data transformations
- **Initialisation scripts were written to deploy all the tools with appropriate orchestration and timing when the Linux VM is launched**
- **Users access and interact with the services via web browser at the frontend after establishing an ssh connection** with a port tunnel specific to the respective service(s)
- **We placed both the Linux and Windows VMs on a daily operating schedule to reduce costs**

Limitations: Some steps related to the maintenance and ingestion of concept mapping document were done manually and will benefit from better version control and a level of automation. Additionally, we are currently ETL-ing the other POS studies in ECRAID-Base to OMOP-CDM and will soon perform some planned studies and analyses with data partners outside and within our consortium to test the scalability, flexibility and ease-of-use of this collaborative, digital research environment-based service architecture.

Discussion:

All infrastructure requirements for ETL validation (ACHILLES, DQD, ARES, ATLAS) were successfully set up.

The data source was published on the EHDEN Database Catalogue ('ECRAID-Base POS VAP').

By starting up the VM's only when needed we were able to reduce our yearly running costs for the two VMs by 95% (i.e. instead of €2450 we ended up spending only €135**).** Although, this is not taking into account the storage accounts as they are continuous costs. Total costs for this setup over the last year were €455.

With environments that require more uptime, costs could be further controlled by scaling the VM to a specific task.

Additionally, **this setup can enable us** to:

- Quickly **increase the number of concurrent users** by simply adding Windows VMs to the network
- Easily **configure the computing power of the backend** to scale up or down, as per task requirements
- **Automate the orchestration of the configuration, deployment and management of the code-based services**
- **Provide and manage the services using predefined templates and automation scripts** to other research teams at UMCU and beyond
- Smoothly **onboard new team members**

Importantly, the ETL will enable us to effectively and efficiently perform studies initiated within the Ecraid consortium and in collaboration with partners in the global OHDSI community and beyond.

Marc Padros Goossens[a], Frank Leus[a], Ben Burke[b], Tom Feusels[b], Jared Houghtaling[b], Freija Descamps[b], Lauren Maxwell[a] and Ankur Krishnan[a]

a European Clinical Research Alliance on Infectious Diseases (Ecraid)
b edenceHealth NV

OHDSI

edence Health

ecraid
Base