AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

LFS303

MELLODDY: Federated machine learning for drug discovery on AWS

Wajahat Aziz
Principal ML/HPC Research SA
Amazon Web Services

Michal Vančo
K8S Cloud Architect
Kubermatic

Wajahat Aziz

Wajahat Aziz

Principal ML/HPC Research SA

- Based in London, United Kingdom
- Principal ML/HPC Specialist Solutions Architect
- Over a decade of experience in HCLS tech with a focus on RWD, clinical trials, and R&D
- Software engineer with 18+ years of full stack development experience
- Working with leading pharma and life sciences companies in helping them define their technology roadmap while leveraging recent advancements in AI/ML and high performance computing

aws

Michal Vančo

Michal Vančo

K8S Cloud Architect, Consultant

- Living in Czech Republic, Brno
- Full-remote consultant in PS team at Kubermatic
- Helping customers with their cloud-native journeys
- WP5 Lead for MELLODDY project
- Full-stack engineer with focus on automation, delivery and QA
- 14 years experience in software engineering

Before . . .

- Full-stack engineer/architect/manager at GoodData
- JBoss Middleware projects at Red Hat

Bringing one drug to market (on average)*

13 years & €1.9 billion

* DiMasi JA *et al.*, 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of Health Economics 47, 20-33.

aws

Competition

Pharma A

Pharma B

Pharma C

Pharma D

MachinE Learning Ledger Orchestration for Drug DiscoverY

MELLODDY

The MELLODDY objectives

MELLODDY aims to show predictive benefits of modelling across tasks, data types, and partners at the largest achievable scale

In three yearly runs, the increasingly sophisticated platform will learn from

- > 10 million annotated small molecules
- > 1 billion assay biological activity labels
- Multiple high-complexity phenotypes at high throughput
- Multiple high-complexity phenotypes at high throughput

Privacy preservation of data and federated models is paramount

aws

Machine learning ledger

ORCHESTRATION FOR DRUG DISCOVERY

MELLODDY

powered by aws

PHARMA PARTNERS

AMGEN

astellas

AstraZeneca

BAYER

Boehringer Ingelheim

gsk

Janssen
PHARMACEUTICAL COMPANIES OF Johnson&Johnson

MERCK

NOVARTIS

SERVIER

PUBLIC PARTNERS

MŰEGYETEM 1782

IKTOS

Kubermatic

KU LEUVEN

NVIDIA

OWKIN

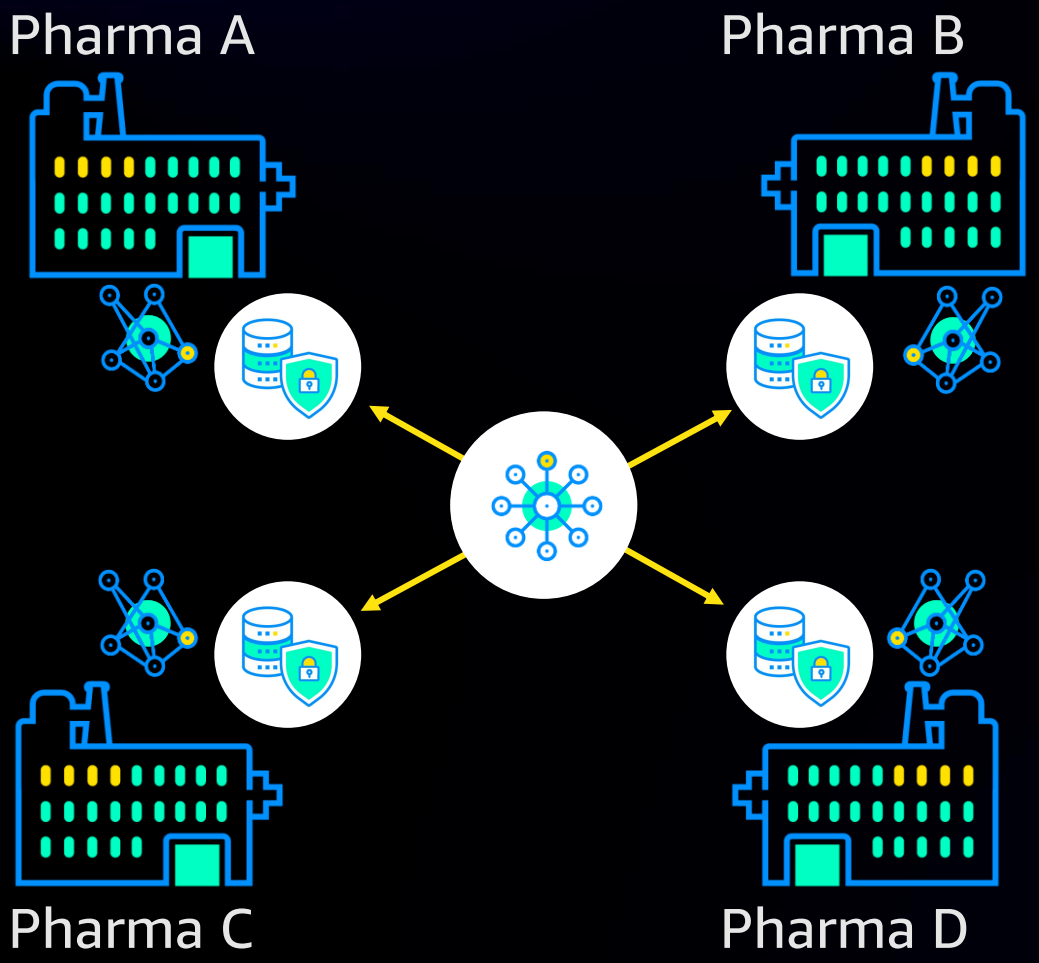
Substra Foundation

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement N° 831472. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA

imi innovative medicines initiative

efpia

Co-opetition



Multitask learning across pharma partners

Compound and activity data and assay-specific models remain under their owner's control

Multi-task approach across partners to improve predictive performance and applicability

AMGEN
astellas
AstraZeneca
BAYER
Boehringer Ingelheim
gsk
janssen
MERCK
NOVARTIS
SERVIER

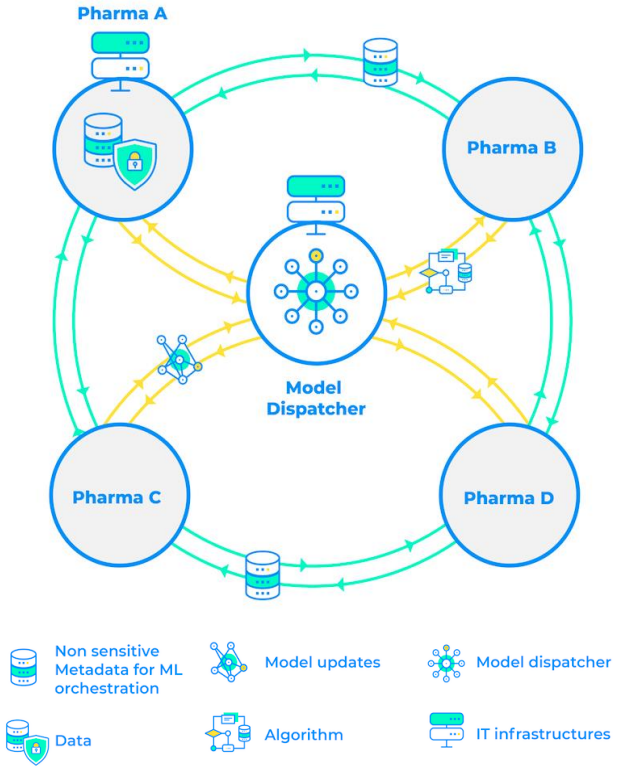
aws

Combined privacy-preserving federated machine learning platform

Sensitive data and assay-specific models remain locked on each pharma's server

Lower level model components are securely exchanged and trained over the network

Complex but transparent pre-agreed access arrangements are strictly enforced



IKTOS

KU LEUVEN

Kubermatic

MŰEGYETEM 1782

NVIDIA

OWKIN

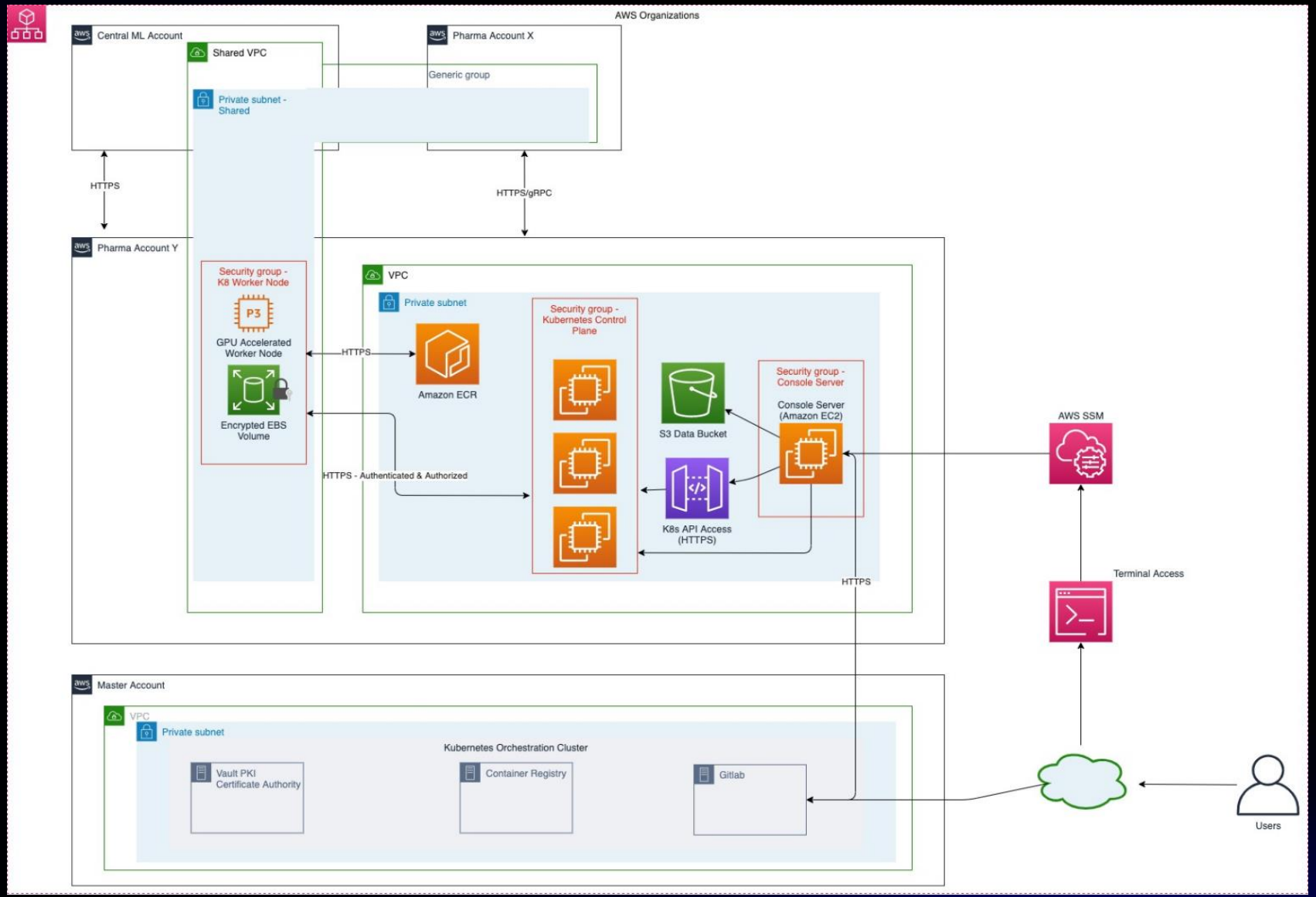
Substra Foundation

Privacy-preserving federated machine learning

Proprietary data and models are hosted in instances owned by each pharma partner

Lower level model components are securely exchanged and trained over the network

Access is restricted and all activities on the platform are transparent



Infrastructure details

KUBERMATIC

Tooling: *KubeOne* vs *Kubermatic Kubernetes Platform*

All AWS resources managed with Terraform modules

Usage of *machine-controller* (Cluster API implementation) to declaratively manage EC2 instances as Kubernetes workers, easy to scale up and down workers

Air-gapped setup

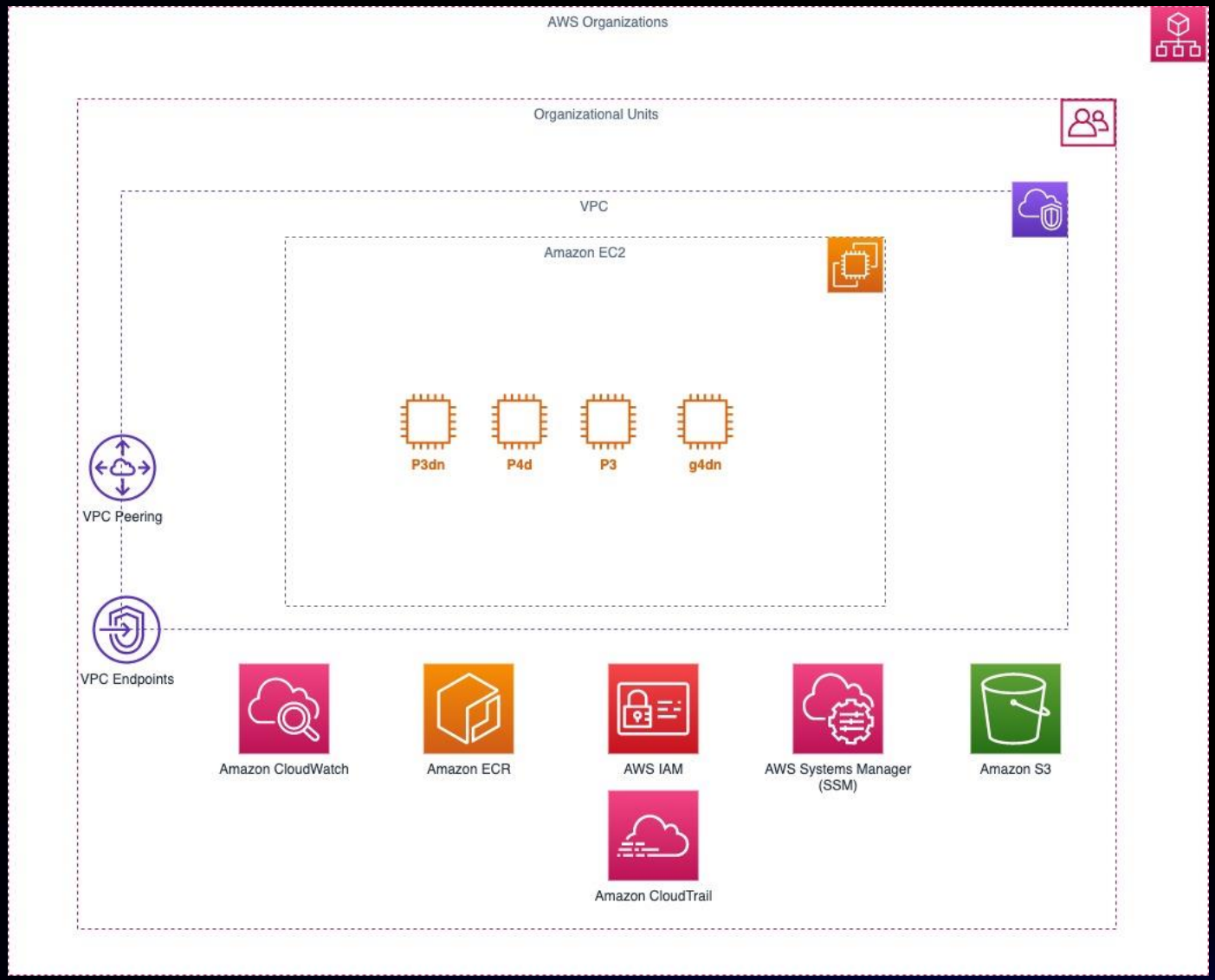
- internet gateway is removed from all nodes due to security reasons
- using custom AMI for Kubernetes workers (based on Ubuntu OS)

Jumphost instance accessed through SSM session, permissions managed via IAM users and roles

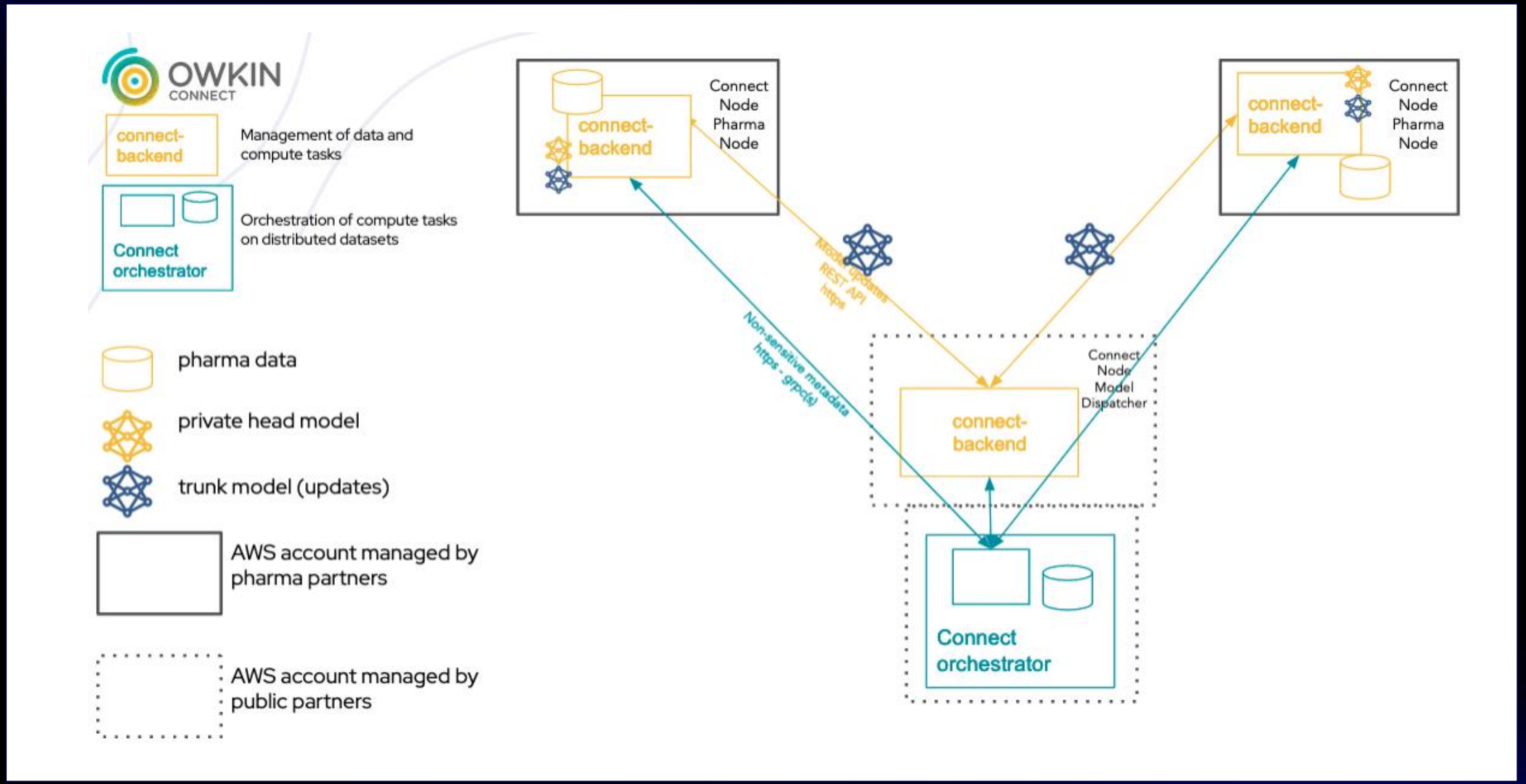
aws

AWS services used

- AWS Organizations – Segregated environment for each pharma partner to curate data and run model training
- Amazon VPC peering and VPC sharing
- Amazon S3 – Artefact storage
- AWS Systems Manager – Configuration management
- IAM - Roles and permissions management
- Amazon ECR – Docker images storage (application/infrastructure)
- Amazon EC2 – g4dn/p3/p4 instances
- CloudWatch, CloudTrail – Monitoring and logging services
- and others

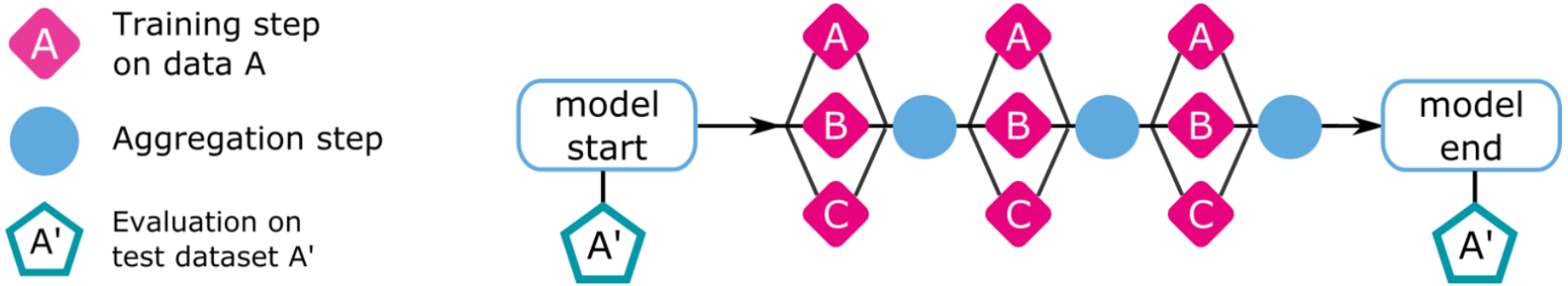


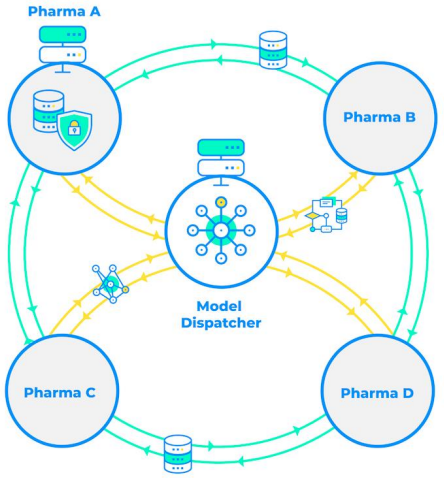
Application details



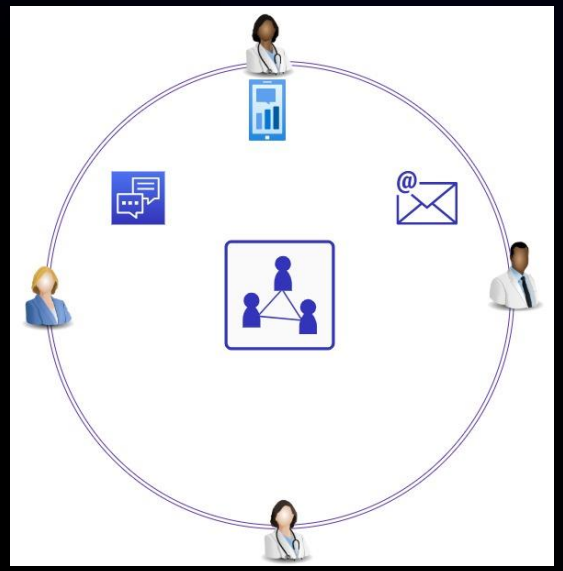
Application workflow

Set of training, aggregation and evaluation tasks



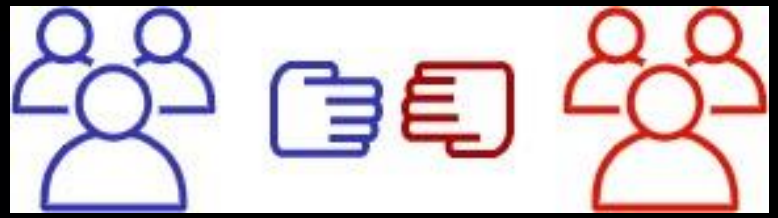


AWS objectives



Better engagement with stakeholders in research across pharma/life sciences

Help customers understand how they can better optimize their AI/ML workloads using our broad spectrum of services



Opportunity for future partnerships involving similar initiatives, specifically concerning privacy-preserving federated machine learning for different healthcare and life sciences use cases

MELLODDY: 3 years, 3 objectives

Year 1: creation of a secure predictive modelling platform, operated at scale

Year 2: study hypothesis that multi-partner modelling yields superior predictive models in drug discovery

Year 3: improve predictive performance

- Secure-by-design
- Audited
- Functional
- Federated training of ML models with pharma partners at scale

FAQs (1/2)

- Do the participants need to use a standard data model or specification to be able to participate in a federated learning run?
- How is data privacy ensured when coordinating model training across accounts?
- How is a new entity or participant onboarded into the platform?
- How do you ensure the security of the platform?
- How did you develop algorithms without seeing the actual data?

FAQs (2/2)

- What types of algorithms can run on the platform?
- How do you ensure the model does not leak data?
- Is there a mechanism for tracing activity on the platform? (e.g. who accessed what assets or artifacts?)
- What further improvements are planned for the platform?

Visit our other life sciences sessions

LFS304 - Workshop: Accelerate science by unifying data silos across the enterprise

LFS302 - Builders Session: Smarter pharmacovigilance with AWS machine learning

Healthcare & Life Sciences Lounge

Join us at our networking lounge everyday for post-session speaker meet-and-greets, ask-the-experts, featured topic deep dives, and informal discussions

---

Wynn, Level 1

aws

Thank you!

Wajahat Aziz
syedazi@amazon.com

Michal Vančo
michal@kubermatic.com