HammerHAI Kick-Off

Prof. Dr. Julian Kunkel
Deputy Head GWDG
Head WG Computing

Prof. Dr. Philipp Wieder
Deputy Head GWDG
Head WG eScience

Dr. Christian Boehme
Deputy Head WG Computing

Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

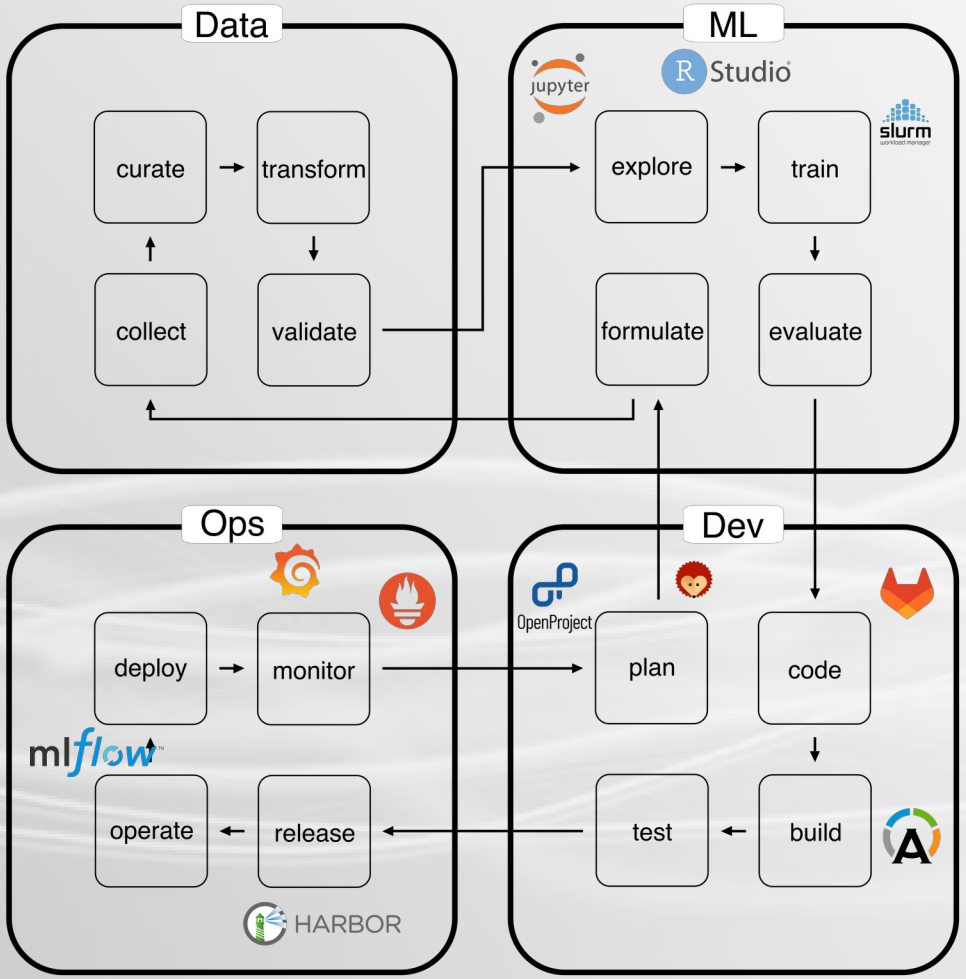
- 05.02.2025

WP 2: OpenStack Cloud Backend

- Main aspects of OpenStack architecture to decide / conceptualize
  - IAM/SSO via Keystone (Backend-IDM? Keycloak?)
  - Network model (SDN vs. VLANs, VPN-access)
  - Partitioning per use case for security?
  - Supported compute platforms (VMs, GPUs via MIG or virtualization?, Bare-Metal?, Kubernetes?)
  - Deployment type and automation (SCS vs. Kolla/Ansible vs. something custom)
  - General operating system choice (aligned with other parts)
  - General security aspects (s. WP4)
- Kubernetes
  - Via OpenStack (Magnum) or
  - Via other frameworks like Rancher (using OpenStack / Bare-Metal-Resources directly)
  - Supported compute platforms (VMs, GPUs via MIG or virtualization, Bare-Metal)

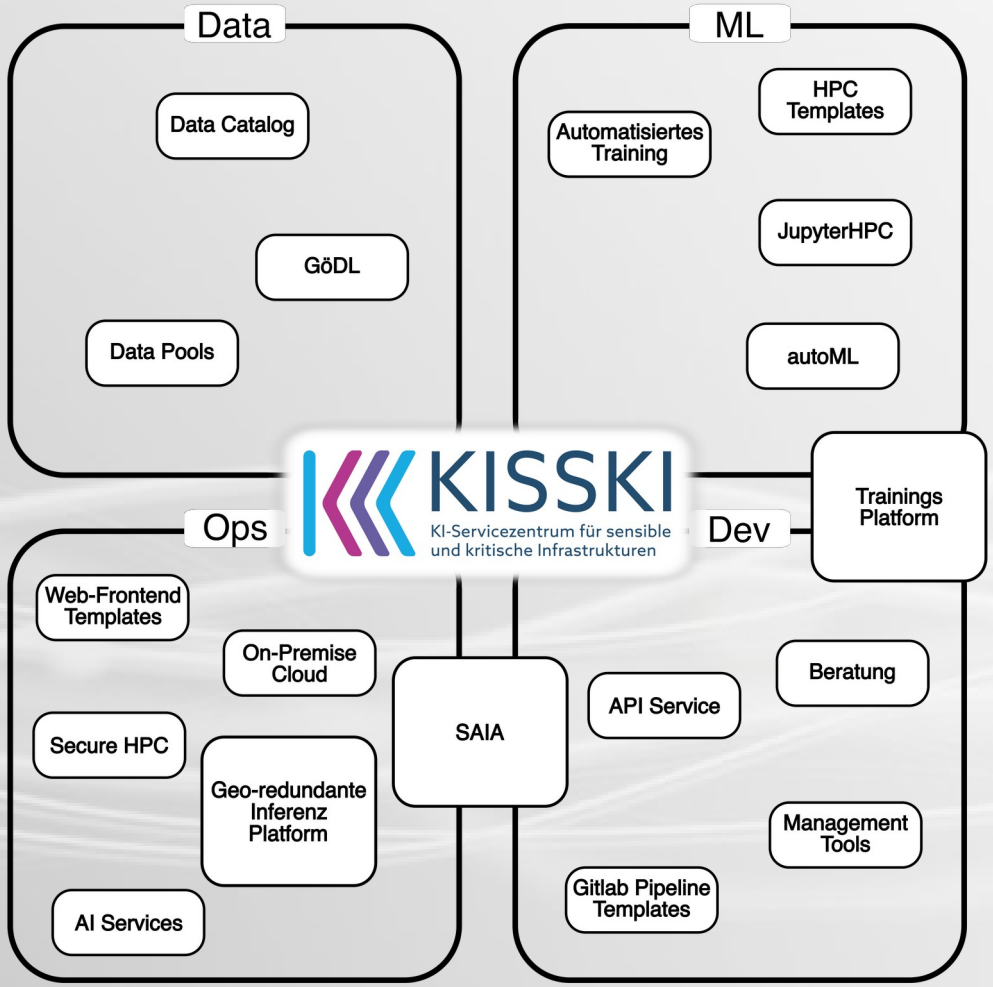
WP 2: MLOps Lifecycle - General

- Data
  - Data Collection and Preprocessing
  - Data Governance and Compliance
- Machine Learning (ML)
  - Model Development and Training Pipelines
  - Model Evaluation and Validation
- Development (Dev)
  - CI/CD Pipelines
  - Code Version Control
- Operations (Ops)
  - Model Deployment and Serving
  - Monitoring and Maintenance



WP 2: MLOps Lifecycle - KISSKI

- Training new models
  - MLOps pipeline in SAIA 2.0
  - Training process via GUI
  - JupyterLab Integration
- Fine-tuning
  - Resource-saving integration of model weights (e.g. LoRa)
  - Fine-tuning web service for models with high demand
- Inference using SAIA 2.0 as broker
  - Deployment of own models in self-service
  - Models can run on third-party systems
  - Simple web frontends from blueprints



WP 4: Secure and Trustworthy Environment

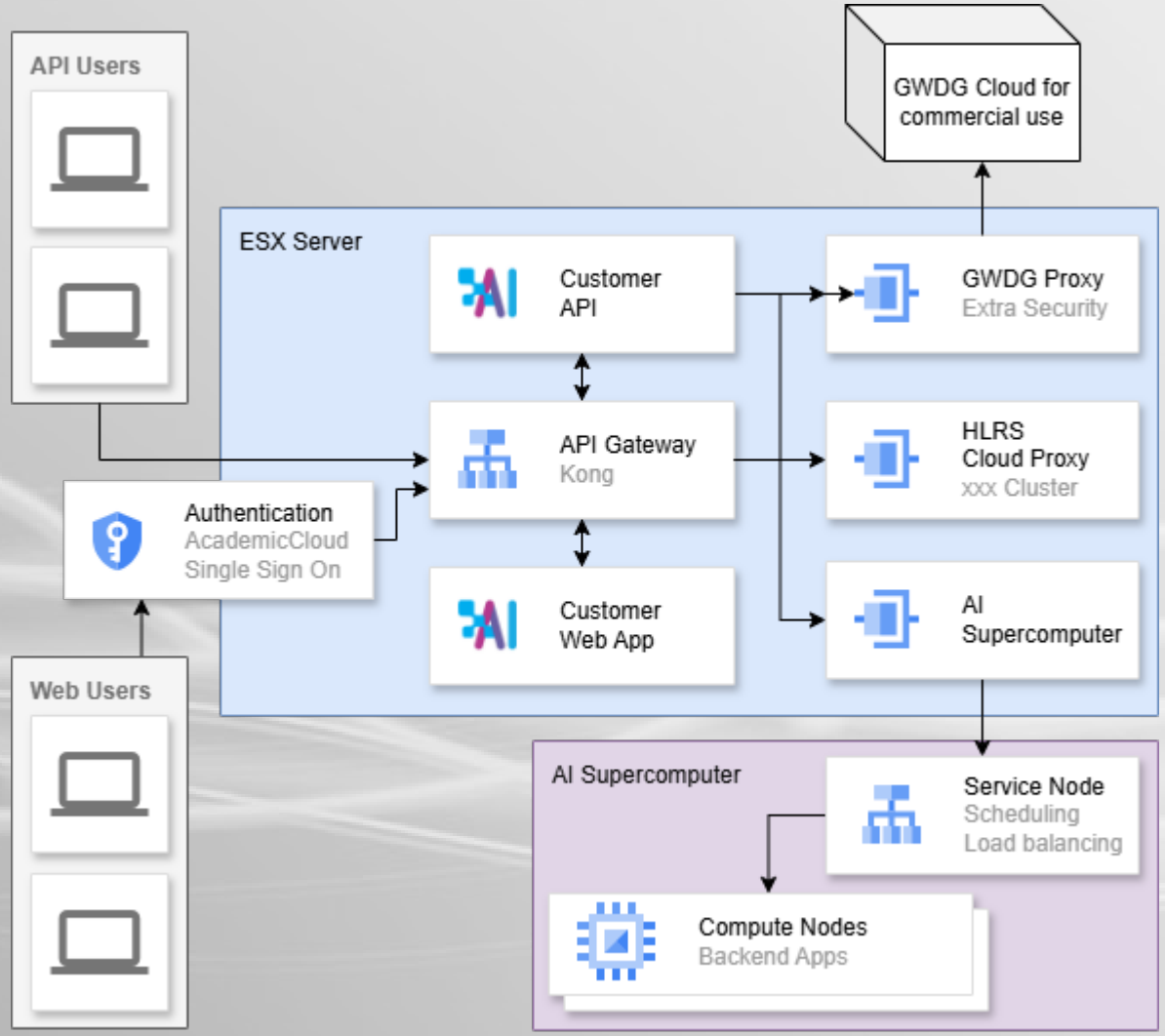
- Participants: **GWDG** (54 PM), USTUTT (54 PM)

- Task 1: Scalable AI Accelerator (SAIA) Platform (**GWDG**, USTUTT, M1-M24)
  - Adapt and extend secure inference platform by GWDG
  - Frontend in hardened GWDG cloud, inference jobs on HPC/cloud
- Task 2: Secure HPC Workflow (**GWDG**, USTUTT, M13-M36)
  - Port workflow by GWDG for highly sensitive data on HPC
  - Encrypted data, containers, batch scripts
- Task 3: Secure Data Transfer and Exchange (**USTUTT**, GWDG, M1-M36)
  - Overcome barriers with data management platforms
  - Integrate USTUTT solutions like SWAN, SCALES

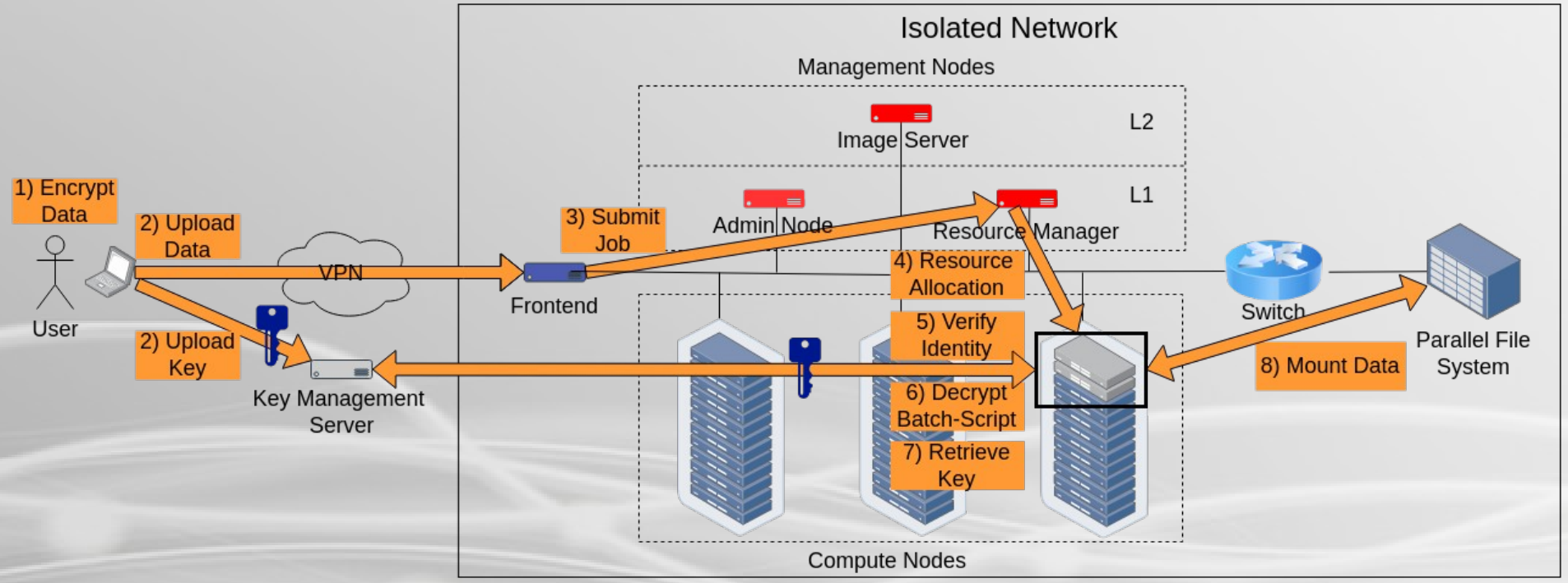
WP 4: Hardened GWDG Cloud

- Identity and Access Management: MFA, role-based access control
- Network Security: VPNs for secure remote access, network segmentation
- Data Encryption: Encrypt data at rest
- Monitoring and Logging: Auditing, analysis, anomaly detection
- Backup and Disaster Recovery
- Security Policies and Training
- ISO 27001 certification
- Security Tools: Intrusion detection, anti-virus, security information and event mgmt.

WP 4: Scalable AI Accelerator (SAIA) Platform



WP 4: Secure HPC Workflow



Secure HPC partition only accessible via dedicated workflow
Data fully end-to-end encrypted (LUKS containers or GoCryptFS)

WP 4: References

- SAIA Platform
  - https://docs.hpc.gwdg.de/services/saia

  - https://github.com/gwdg/saia-hpc
  - https://github.com/gwdg/saia-hub
  - https://arxiv.org/abs/2407.00110

- Secure HPC Workflow
  - https://docs.hpc.gwdg.de/services/secure-hpc

  - https://github.com/gwdg/secure-hpc
  - https://doi.org/10.1016/j.future.2022.12.019