PRODYNA STACKIT
A Brand of Schwarz Digits

Agentic AI Platform on STACKIT with Kubernetes

Motivation

For most organizations, the challenge with generative and agentic AI is no longer the prototyping stage, but rather operating the technology reliably on a large scale while retaining data sovereignty. Credentials are duplicated across teams, model consumption cannot be attributed to its source, each agent integrates their own tools and deployment practices diverge from project to project. Where regulation or sovereignty requirements apply, routing sensitive prompts through external LLM APIs raises further concerns.

This offering establishes a governed platform built on STACKIT as a sovereign cloud, ensuring that data and AI workloads remain under German jurisdiction. A central AI gateway provides a single point of access to all models (STACKIT-managed or self-hosted) and to the governed MCP tool servers. This means that access control, budgets and guardrails can be enforced centrally rather than on a project-by-project basis.

What We Bring

With more than two decades of experience in enterprise software development, cloud-native architecture, and Kubernetes operations, PRODYNA is well-placed to support your business. We deliver a governed, scalable Agentic AI platform built on standard, enterprise-grade open-source technologies.

- A central open-source AI gateway (AgentGateway) for unified, OpenAI-compatible access to STACKIT-managed and self-hosted models with single sign-on, RBAC at group level, credential injection, token budgets, guardrails, and cost attribution at agent level
- A governed MCP gateway that fronts Model Context Protocol tool servers with shared authentication, per-server authorisation, and a self-service registry so that agents only discover the tools they are permitted to use
- A ready-to-use chat UI (LibreChat) with single sign-on, per-user model menus, token budgets and MCP tools in chat
- A GitOps-based delivery workflow (CI/CD, hardened container images and ArgoCD) for zero-downtime deployments
- Built-in observability with the Grafana stack: Prometheus metrics, Grafana Tempo (optional Langfuse) for distributed LLM & agent tracing, and ready made dashboards for usage, cost and budgets
- Optional self-hosted LLMs on Kubernetes (vLLM) can be managed optionally by KServe
- Hands-on enablement: we develop the initial AI agents alongside your team

What You Need

To make the best use of this offer and permit a fast and efficient start, you will need:

- A STACKIT Organization + Foundation (Landing Zones)
- Availability of your experts (e.g. platform, network, IAM, security)
- A small group of motivated IT colleagues to onboard as early adopters

Benefits

- **Sovereign by design:** Runs on Kubernetes on STACKIT, keeping data and AI workloads under German jurisdiction
- **Predictable costs:** Self-host open-weight models on your own GPU capacity instead of paying dynamic per-token prices
- **Self-service for developers:** An opinionated Golden Path template ships AI applications in days, without deep Kubernetes expertise
- **Governed from day one:** Gateway, SSO and RBAC, cost tracking, guardrails and observability are inherited by every agent
- **Proven foundation:** Standard enterprise open-source with hands-on enablement

Quick facts

- **Duration:** ~10-12 weeks, depending on use case and cluster setup
- **Deliverables:** Platform with gateway and observability, first AI agents, team enablement

PRODYNA

What You Get

PHASE	ACTIONS	DELIVERABLES
Kubernetes Setup (2 weeks)	• Provision a production-ready Kubernetes cluster (control plane, node pools, optional GPU nodes) • Configure ingress/gateway, secret management (vault), storage, and network policies • Establish baseline platform security and isolation model	• Hardened Kubernetes cluster ready to host the AI platform • Reusable Terraform modules for cluster operations
AI Platform (MVP) (4-6 weeks)	• Deploy SSO and identity (OIDC): JWT authentication and per-group RBAC • Deploy the LLM Gateway (agentgateway): unified model access, upstream-credential injection, token budgets, guardrails, and per-user/per-agent cost tracking • Deploy the MCP gateway: governed Model Context Protocol tool servers with per-server authorization and a self-service registry • Deploy the chat UI (LibreChat) with SSO, per-user model menus and MCP tools in chat • Deploy observability: Grafana Tempo (LLM & agent tracing) with the Prometheus/Grafana stack (metrics, dashboards & logs) • Optional: self-hosted open-weight LLMs on GPU nodes with e.g. vLLM/SGLang (optionally via KServe)	• Running Agentic AI Platform with governed model and tool access (SSO + RBAC) • Central cost control, token budgets and full observability
GitOps Agent Deployment (4 weeks)	• Implement the "Golden Path" agent template (OpenAI-compatible service, agent.yaml) — every agent is automatically fronted by the gateway and inherits SSO, RBAC, budgets, guardrails and tracing • Set up CI/CD pipelines, container registry, and ArgoCD GitOps sync • Build the first agents and MCP tool servers side-by-side with early adopters • Knowledge transfer and onboarding documentation	• Reproducible self-service onboarding path • First AI Agent running • Enabled internal team able to onboard further use cases

Note: This is our standard methodology. We are happy to modify and adapt to the specific needs of your organization.

Get Started

To learn about pricing and how to get started, please contact info@prodyna.com.

About PRODYNA

PRODYNA is a technology consulting and software engineering company that helps organizations create business value through AI, cloud, data, and custom digital platforms, delivering solutions from strategy through implementation and operation.