Patryk Kozak

Senior AI Platform Engineer · AI Solution Engineer · MLOps & LLMOps

patrykpiotrkozak@gmail.com · +43 660 2656336 · Wien, Österreich · LinkedIn · GitHub

PROFIL

AI Platform Engineer mit über drei Jahren Projekterfahrung bei Hewlett Packard Enterprise (Advisory & Professional Services, DACH) im Aufbau und Betrieb von Infrastruktur für Generative-AI- und ML-Workloads auf Kubernetes – On-Premises auf HPE Private Cloud AI und HPE Ezmeral, mit Anbindung AWS-basierter Systeme und Datenquellen. Schwerpunkte: Observability und Betriebsfähigkeit von LLM-Inference-Stacks (NVIDIA NIM, vLLM, Milvus, GPU), LLMOps mit Guardrails und Evaluierung (NVIDIA NeMo Guardrails, MLflow), CI/CD- und GitOps-Automatisierung für Kubernetes (GitHub Actions, Jenkins, Argo CD, Helm) sowie Python-basierte ML-Services und Data Engineering. Ich verbinde hands-on Engineering mit der Abstimmung zwischen Data-Science-, Applikations- und Betriebsteams.

TECHNISCHE KOMPETENZEN

AI-Plattform & Infrastruktur: Kubernetes (Multi-Tenant), Helm, Docker, Linux, HPE Private Cloud AI, HPE Ezmeral Runtime Environment, HPE Machine Learning Development Environment, NVIDIA NIM, vLLM, NVIDIA GPU / CUDA, Milvus

MLOps & LLMOps: MLflow (Experiment-Tracking, LLM-Evaluierung), Kubeflow, Apache Airflow, NVIDIA NeMo Guardrails, RAG-Pipelines und Retrieval-Optimierung, LangChain, LangGraph, LlamaIndex, Model Context Protocol (MCP), OpenAI Agents SDK, LLM-Observability

CI/CD & GitOps: GitHub Actions, Jenkins, Argo CD, Helm Charts, Git-basierte Deployment-Workflows

Observability & Betrieb: Prometheus, PromQL, Alertmanager, Grafana, Loki, Promtail, Grafana Alloy, NVIDIA DCGM Exporter, SLO-Monitoring

Security & Zugriff: TLS-Zertifikatsmanagement, Kubernetes RBAC, Secrets-Management, Network Policies, mandantenfähige Berechtigungen in Grafana

Cloud: AWS (Amazon S3, Integration AWS-basierter Systeme und Datenquellen)

Software Engineering: Python (FastAPI, pandas, NumPy, pytest), REST-APIs, Git, Docker, Code Reviews

Data Engineering: Apache Kafka, Apache Flink, Apache Airflow, PostgreSQL, Neo4j / Cypher

Data Science & ML: PyTorch, TensorFlow, NVIDIA RAPIDS (cuDF, cuML), Time-Series Forecasting, Neo4j Graph Data Science

BERUFLICHER WERDEGANG

Data Scientist · Technology Consultant, *Hewlett Packard Enterprise*
Wien, Österreich (Hybrid)
02/2023 – 06/2026

Advisory & Professional Services – DACH Delivery. Hands-on Umsetzung von Kundenprojekten in den Bereichen AI-Infrastruktur, MLOps/LLMOps, Plattform-Engineering und Data Engineering für Automobilindustrie, öffentlichen Sektor, Bauwesen, industrielle Distribution und Sportorganisationen.

GenAI-Plattform & LLMOps-Observability – Digitaler KI-Avatar, staatliches Immobilienunternehmen

KI-Avatar mit RAG und LLM-Inference (NVIDIA NIM, vLLM, Milvus) auf einem Multi-Tenant-Kubernetes-Cluster (On-Premises). Rolle: Aufbau und Betrieb der Observability, Mitarbeit an LLM-Qualität, RAG und Agentic AI.

- Aufbau eines End-to-End-Observability-Stacks von Grund auf (Prometheus, Alertmanager, Grafana, Loki, Promtail / Grafana Alloy, NVIDIA DCGM Exporter), betrieben direkt auf dem Kubernetes-Cluster über alle Namespaces; Deployment per Helm Charts, Konfiguration versioniert in Git.
- Metrikerfassung per Prometheus Scraping der Service-Ports und /metrics-Endpunkte, Weiterleitung der Container-Logs an Loki; Instrumentierung des gesamten Inference-Stacks: Inferenzlatenz (inkl. Time-to-First-Token), Token-Durchsatz, Fehlerraten, GPU-Auslastung und -Speicher (DCGM), Milvus- und RAG-Pipeline-Metriken.

- Alerting-Regeln in Prometheus (PromQL) mit automatischen Benachrichtigungen über Alertmanager; Grafana-Dashboards und SLO-Tracking, über die das Team Latenz, Durchsatz, GPU-Auslastung und Fehlerraten der kundenorientierten GenAI-Workloads zentral überwachte.
- LLM-Qualität und -Sicherheit: Absicherung der LLM-Antworten mit NVIDIA NeMo Guardrails (Themen- und Antwortgrenzen); Evaluierung von Modell-, Prompt- und RAG-Varianten mit MLflow als Grundlage für Release-Entscheidungen.
- Optimierung der semantischen Suche und des RAG-Retrievals in Milvus (Embeddings, Index- und Suchparameter, Top-k) zur Verbesserung der Antwortqualität.
- Mandantenfähige Sicht auf Metriken und Logs über Namespaces hinweg (Label-Konzept, Grafana-Ordner und -Berechtigungen); Einbindung der Observability-Komponenten in das im Team umgesetzte Sicherheitskonzept des Clusters (Kubernetes RBAC, Secrets-Management, Network Policies).
- Performance- und Kapazitätsanalysen: Identifikation von Bottlenecks und Kapazitätsengpässen im Inference-Stack; Ableitung von Empfehlungen zu GPU-Zuteilung, Replikas und Skalierung.
- Mitentwicklung, Containerisierung und Integration von Agentic-AI- und RAG-Komponenten (LangChain / LangGraph, Tool-Integration, Retrieval aus Milvus) in zuverlässige Produktions-Workflows.

Platform Engineering – HPE Ezmeral Runtime Environment, globaler Premium-Automobilhersteller

Kubernetes-basierte Container-Plattform (HPE Ezmeral Runtime Environment) für Analytics- und AI-Workloads im Produktionsbetrieb.

- Entwicklung, Automatisierung und Validierung einer Zero-Downtime-Strategie für die Rotation der Plattform-Zertifikate – Test in Nicht-Produktivumgebung, Runbook und Rollback-Pfad, anschließend Umsetzung in Produktion.
- Post-Installation-Konfiguration und Stabilisierung der Plattform (Identity-Anbindung, Storage, Health-Checks) zur Unterstützung zuverlässiger Produktionsbereitstellungen.

CI/CD & GitOps – Kubernetes-Deployments (projektübergreifend)

- Aufbau und Pflege von CI/CD-Pipelines mit GitHub Actions und Jenkins: automatisierte Tests, Container-Image-Builds und Deployments auf Kubernetes.
- Paketierung von Services und Plattform-Komponenten als Helm Charts mit umgebungsspezifischer Konfiguration (Values pro Umgebung).
- GitOps-basierte Deployments mit Argo CD: deklarativer Soll-Zustand in Git, automatische Synchronisation in den Cluster, nachvollziehbare Änderungen und Rollbacks über die Git-Historie.

MLOps & Time-Series Forecasting – Internes HPE-Produktprojekt (HPE MLDE, AWS-basiert)

- Mitentwicklung eines Forecasting-Microservice in Python über den gesamten ML-Lifecycle: Data Ingestion (Batch & Streaming), Training, Deployment und Inference.
- Ausführung der ML-Workloads (Training, Experiment-Tracking, Modellversionierung) in der AWS-basierten HPE Machine Learning Development Environment (Determined AI); Anbindung AWS-basierter Datenquellen (Amazon S3).
- REST-API (FastAPI) zur Bereitstellung der Forecasts für HPE Private Cloud AI, LLM-basierte Anwendungen und Enterprise-Systeme; containerisierte Bereitstellung auf Kubernetes.
- Unit- und Integrationstests (pytest), Code Reviews im Git-Workflow.
- Mitarbeit am Time-Series- und Stream-Processing-Architekturdesign mit Apache Flink.

Data Engineering & Graph Analytics – Globaler Automobilhersteller

- Entwicklung einer Python-Data-Transformation-Engine, die Daten über Apache Kafka in strukturierte Analytics-Pipelines (PostgreSQL, Neo4j) integriert; Containerisierung der Services mit Docker für produktive Deployments.
- Co-Design und Entwicklung eines Python-basierten Similarity-Search-Algorithmus mit kundenspezifischen Evaluierungsmetriken (Jaccard, Pearson, Spearman) zur Verbesserung der Metadaten-Retrieval-Qualität.
- Neo4j Graph Data Science (Similarity, Centrality, Path-Finding) und Supervised-Classification-Workflows auf Metadaten; Verwaltung der PostgreSQL- und Neo4j-Datenbanken; EDA und Validierung der ML-Ergebnisse.

Data Engineering – Professionelle Sportorganisation

- Entwicklung und Scheduling von Apache-Airflow-DAGs für die Datenpipelines einer Lottery-basierten Fan-Engagement-Plattform; zuverlässige, auditierbare Verarbeitung mit Retries und Monitoring.

Python Automation & Infrastructure Support – Internationales Handels- und Distributionsunternehmen

- Python-Automatisierung für File Retention, Log Rotation und serverseitiges Data Lifecycle Management (Scheduled Jobs); Unterstützung bei Server-Upgrades, Zertifikatserneuerungen und Runtime-Migrationen.

Data-Centric Architecture – Nationales Bauunternehmen

- Workshops nach dem HPE Edge-to-Cloud Adoption Framework (E2CAF): Mapping von Datenfähigkeiten, Zugriffskonzepten und RBAC-Modellen; Unterstützung beim Aufbau eines Transformation Business Office für Governance und Entscheidungsfindung.

Weiterbildung, Community & Thought Leadership

- HPE & NVIDIA AI Development Initiative (AI-Infrastruktur-Betrieb, GenAI, LLMs, multimodale AI); technische Trainings zu HPE Private Cloud AI und Azure ML (MLOps Deployment Patterns).
- LLM-Hackathon: Design und Entwicklung eines autonomen AI-Agents mit LangChain; technische Präsentationen zu AI-basierter Luftbildanalyse und zum HPE-GenAI-Portfolio; HPE DeliverAI Sessions.

Google Ads Global Care Associate, *Google (via Accenture)*

Warschau, Polen (Remote, Dublin-Team) · 08/2021 – 10/2022

- Betreuung und Analyse von Google-Ads-Kampagnen für Werbetreibende im DACH-Markt; Diagnose und Behebung von Kampagnenproblemen.

Data Analyst, *Accenture*

Warschau, Polen · 05/2020 – 07/2021

- Analyse und Qualitätssicherung großer Social-Media-Datensätze (Korrelationsanalysen, Visualisierung, Anomalie- und Duplikaterkennung).

B2B Sales Specialist, *Colligo GmbH*

Wien, Österreich · 09/2018 – 07/2019

- B2B-Vertrieb und Digital-Marketing-Services für Kunden in der DACH-Region.

AUSBILDUNG

Software Development Academy, *Certified Data Scientist – Professional Diploma*
Warschau, Polen · 09/2020 – 07/2022 · Zweijähriges Data-Science-Programm (Python, Statistik, Machine Learning, SQL, Datenvisualisierung, Projektarbeiten)

365 Data Science, *Certified Data Scientist – Professional Career Diploma*
Certificate ID: DD-D9D0C2C28B · Wahrscheinlichkeitsrechnung, Statistik, Python, SQL, Machine Learning, Deep Learning

Dr. Roland AHS, *Matura*
Wien, Österreich · 06/2016

ZERTIFIZIERUNGEN & WEITERBILDUNGEN

NVIDIA-Certified Associate: AI Infrastructure & Operations (NCA-AIIO); Generative AI with LLMs (NCA-GENL); Generative AI Multimodal (NCA-GENM), *NVIDIA*

NVIDIA DLI: AI Infrastructure & Operations Fundamentals; Transformer-Based NLP; Building AI Agents with Multimodal Models; Building Conversational AI Applications; Accelerating End-to-End Data Science Workflows, *NVIDIA Deep Learning Institute*

HPE NVIDIA AI Development Initiative, *HPE & NVIDIA*

LFS207: Linux System Administration Essentials; Docker & Kubernetes (self-paced), *The Linux Foundation*

Machine Learning Specialization; Deep Learning Specialization, *DeepLearning.AI*

Neo4j Graph Data Science; Neo4j Cypher Query Language, *Neo4j*

Microsoft Azure Data Science Track, *Microsoft* · **Software Engineering for Data Scientists in Python**, *DataCamp* · **Python Developer**, *ZTM Academy* · **Time Series Modeling**, *Skillsoft* · **Agile Project Management**, *Accenture*

SPRACHEN

Deutsch (zweisprachig) · **Englisch** (verhandlungssicher) · **Polnisch** (Muttersprache)