Patryk Kozak

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

patrykpiotrkozak@gmail.com · +43 660 2656336 · Vienna, Austria · LinkedIn · GitHub

PROFILE

AI Platform Engineer with over three years of project experience at Hewlett Packard Enterprise (Advisory & Professional Services, DACH), building and operating infrastructure for generative AI and ML workloads on Kubernetes - on-premises on HPE Private Cloud AI and HPE Ezmeral, integrating AWS-based systems and data sources. Focus areas: observability and operability of LLM inference stacks (NVIDIA NIM, vLLM, Milvus, GPU), LLMOps with guardrails and evaluation (NVIDIA NeMo Guardrails, MLflow), CI/CD and GitOps automation for Kubernetes (GitHub Actions, Jenkins, Argo CD, Helm), and Python-based ML services and data engineering. I combine hands-on engineering with coordination across data science, application, and operations teams.

TECHNICAL SKILLS

AI Platform & Infrastructure: 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 evaluation), Kubeflow, Apache Airflow, NVIDIA NeMo Guardrails, RAG pipelines and retrieval optimization, LangChain, LangGraph, LlamaIndex, Model Context Protocol (MCP), OpenAI Agents SDK, LLM observability

CI/CD & GitOps: GitHub Actions, Jenkins, Argo CD, Helm Charts, Git-based deployment workflows

Observability & Operations: Prometheus, PromQL, Alertmanager, Grafana, Loki, Promtail, Grafana Alloy, NVIDIA DCGM Exporter, SLO monitoring

Security & Access: TLS certificate management, Kubernetes RBAC, secrets management, Network Policies, multi-tenant permissions in Grafana

Cloud: AWS (Amazon S3, integration of AWS-based systems and data sources)

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

PROFESSIONAL EXPERIENCE

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

Advisory & Professional Services - DACH Delivery. Hands-on implementation of client projects in AI infrastructure, MLOps/LLMOps, platform engineering, and data engineering for the automotive industry, public sector, construction, industrial distribution, and sports organizations.

GenAI Platform & LLMOps Observability - Digital AI Avatar, State-Owned Real Estate Company

AI avatar with RAG and LLM inference (NVIDIA NIM, vLLM, Milvus) on a multi-tenant Kubernetes cluster (on-premises). Role: building and operating observability; contributing to LLM quality, RAG, and agentic AI.

- Built an end-to-end observability stack from the ground up (Prometheus, Alertmanager, Grafana, Loki, Promtail / Grafana Alloy, NVIDIA DCGM Exporter), running directly on the Kubernetes cluster across all namespaces; deployed via Helm Charts, with configuration versioned in Git.
- Collected metrics through Prometheus scraping of service ports and /metrics endpoints, and forwarded container logs to Loki; instrumented the entire inference stack: inference latency (including time to first token), token throughput, error rates, GPU utilization and memory (DCGM), and Milvus and RAG pipeline metrics.

- Configured alerting rules in Prometheus (PromQL) with automatic notifications through Alertmanager; created Grafana dashboards and SLO tracking that enabled the team to centrally monitor latency, throughput, GPU utilization, and error rates for customer-facing GenAI workloads.
- LLM quality and safety: safeguarded LLM responses with NVIDIA NeMo Guardrails (topic and response constraints); evaluated model, prompt, and RAG variants with MLflow to inform release decisions.
- Optimized semantic search and RAG retrieval in Milvus (embeddings, index and search parameters, top-k) to improve response quality.
- Provided multi-tenant visibility into metrics and logs across namespaces (labeling scheme, Grafana folders and permissions); integrated observability components into the cluster security framework implemented by the team (Kubernetes RBAC, secrets management, Network Policies).
- Performance and capacity analysis: identified bottlenecks and capacity constraints in the inference stack; developed recommendations for GPU allocation, replicas, and scaling.
- Co-developed, containerized, and integrated agentic AI and RAG components (LangChain / LangGraph, tool integration, retrieval from Milvus) into reliable production workflows.

Platform Engineering - HPE Ezmeral Runtime Environment, Global Premium Automotive Manufacturer

Kubernetes-based container platform (HPE Ezmeral Runtime Environment) for analytics and AI workloads in production.

- Developed, automated, and validated a zero-downtime strategy for rotating platform certificates - testing in a non-production environment, preparing a runbook and rollback procedure, then implementing it in production.
- Configured and stabilized the platform after installation (identity integration, storage, health checks) to support reliable production deployments.

CI/CD & GitOps - Kubernetes Deployments (Across Projects)

- Built and maintained CI/CD pipelines with GitHub Actions and Jenkins: automated testing, container image builds, and deployments to Kubernetes.
- Packaged services and platform components as Helm Charts with environment-specific configuration (values per environment).
- Implemented GitOps-based deployments with Argo CD: declarative desired state in Git, automatic synchronization to the cluster, and traceable changes and rollbacks through Git history.

MLOps & Time-Series Forecasting - Internal HPE Product Project (HPE MLDE, AWS-Based)

- Co-developed a forecasting microservice in Python across the entire ML lifecycle: data ingestion (batch and streaming), training, deployment, and inference.
- Ran ML workloads (training, experiment tracking, model versioning) in the AWS-based HPE Machine Learning Development Environment (Determined AI); integrated AWS-based data sources (Amazon S3).
- Provided forecasts through a REST API (FastAPI) for HPE Private Cloud AI, LLM-based applications, and enterprise systems; deployed the containerized service on Kubernetes.
- Conducted unit and integration tests (pytest) and code reviews within the Git workflow.
- Contributed to the design of the time-series and stream-processing architecture with Apache Flink.

Data Engineering & Graph Analytics - Global Automotive Manufacturer

- Developed a Python data transformation engine that integrates data via Apache Kafka into structured analytics pipelines (PostgreSQL, Neo4j); containerized services with Docker for production deployments.
- Co-designed and developed a Python-based similarity search algorithm with client-specific evaluation metrics (Jaccard, Pearson, Spearman) to improve metadata retrieval quality.
- Applied Neo4j Graph Data Science (similarity, centrality, pathfinding) and supervised classification workflows to metadata; administered PostgreSQL and Neo4j databases; performed EDA and validated ML results.

Data Engineering - Professional Sports Organization

- Developed and scheduled Apache Airflow DAGs for the data pipelines of a lottery-based fan engagement platform; supported reliable, auditable processing with retries and monitoring.

Python Automation & Infrastructure Support - International Trading and Distribution Company

- Developed Python automation for file retention, log rotation, and server-side data lifecycle management (scheduled jobs); supported server upgrades, certificate renewals, and runtime migrations.

Data-Centric Architecture - National Construction Company

- Conducted workshops using the HPE Edge-to-Cloud Adoption Framework (E2CAF): mapped data capabilities, access models, and RBAC models; supported the establishment of a Transformation Business Office for governance and decision-making.

Professional Development, Community & Thought Leadership

- HPE & NVIDIA AI Development Initiative (AI infrastructure operations, GenAI, LLMs, multimodal AI); technical training in HPE Private Cloud AI and Azure ML (MLOps deployment patterns).
- LLM hackathon: designed and developed an autonomous AI agent with LangChain; technical presentations on AI-based aerial image analysis and the HPE GenAI portfolio; HPE DeliverAI Sessions.

Google Ads Global Care Associate, *Google (via Accenture)*
Warsaw, Poland (Remote, Dublin Team) · 08/2021 - 10/2022

- Supported and analyzed Google Ads campaigns for advertisers in the DACH market; diagnosed and resolved campaign issues.

Data Analyst, *Accenture*
Warsaw, Poland · 05/2020 - 07/2021

- Analyzed and performed quality assurance on large social media datasets (correlation analysis, visualization, anomaly and duplicate detection).

B2B Sales Specialist, *Colligo GmbH*
Vienna, Austria · 09/2018 - 07/2019

- B2B sales and digital marketing services for clients in the DACH region.

EDUCATION

Software Development Academy, *Certified Data Scientist – Professional Diploma*
Warsaw, Poland · 09/2020 - 07/2022 · Two-year data science program (Python, statistics, machine learning, SQL, data visualization, project work)

365 Data Science, *Certified Data Scientist – Professional Career Diploma*
Certificate ID: DD-D9D0C2C28B · Probability, statistics, Python, SQL, machine learning, deep learning

Dr. Roland AHS, *Matura (upper secondary school-leaving qualification)*
Vienna, Austria · 06/2016

CERTIFICATIONS & PROFESSIONAL DEVELOPMENT

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*

LANGUAGES

German (bilingual) · **English** (business fluent) · **Polish** (native)