Nikolay Penkov

Phone: +359893991263
Mail: n.penkow@gmail.com
Website: https://penkow.com
LinkedIn: @penkow
Medium: @penkow

Experienced Cloud Engineer
passionate about building cutting-edge cloud infrastructure and software solutions.

EXPERIENCE

Cloud/MLOps Engineer · Mentalyc Inc
01/2024 - 01/2025
San Francisco, USA (remote)

- Developed an in-house ML based audio transcription service leveraging state-of-the-art AI models. Deployed the workload using AWS SageMaker with a custom containerized setup. Later scaled it with Kubernetes and GPU virtualization for better cost and performance.
- Achieved 10x cost reduction and 3x processing speed improvement, boosting client acquisition and retention.
- Lead an infrastructure team to extend company processes and cloud infrastructure for SOC2 Type II compliance.

Technologies: AWS (VPC, EC2, ECS, EKS, SM, IAM, DynamoDB, Route 53, RDS, S3), Kubernetes, Terraform, NVIDIA GPU Operator, Vanta, PagerDuty, NodeJS, Python, OpenAI API, HuggingFace

Machine Learning Platform Engineer · SAP
09/2023 - 04/2024
Sofia, Bulgaria (on-site)

- Worked on improving and expanding an internal MLOps platform, enabling faster experimentation and efficient model deployment.
- Collaborated with the PO and senior DevSecOps engineers to extend and scale the cloud infrastructure in a compliant manner.
- Contributed to the hiring process by conducting technical interviews for the expansion of the MLOps team.

Technologies: Azure (Entra ID, DataBricks, Virtual Networks, Virtual Machines, Blob Storage), Docker, Python, Jenkins, Terraform

Certifications

Certified Solutions Architect (**SAA-C03**)

Certified Kubernetes Application Developer (**CKAD**)

Achievements

Differentiable Slimming for Memory-Efficient Transformers

SWaTEval: An Evaluation Framework for Stateful Web Application Testing

Education

Mechatronics and AI
KIT, Germany

Master of Science specializing in system optimization, machine learning, statistical methods, and artificial intelligence (AI).

Languages

English
Full Professional Proficiency

German
Full Professional Proficiency

Bulgarian
Native Language

Cloud/MLOps Engineer · esentri AG

01/2023 - 08/2023

Karlsruhe, Germany (on-site)

- Worked on a showcase project presenting the capabilities of end-to-end ML systems in the industrial analytics domain.
- Developed a ML model for time-series classification of IoT sensor vibration data for continuous learning.
- Deployed the setup to production leveraging serverless architecture for minimal infrastructure cost.

Technologies: AWS (Lambda, Kinesis, DynamoDB, GreenGrass), PyTorch, Python, Jupyter, Terraform

Graduate Student Researcher · Fraunhofer IOSB, CES KIT

05/2021 - 01/2023

Karlsruhe, Germany (on-site)

- Conducted advanced research on Cyber Security and Large Language Optimization topics.
- Developed a novel method to identify and prune redundant model components post-training, reducing complexity while maintaining performance.
- Also contributed to the development of a state-aware security vulnerability scanner using AI and ML techniques.

Expertise: Scientific Research, Technical Writing, Linux, OWASP Top 10, Machine Learning, Data Science

Founder and Technical Lead · Own Startup

02/2020 - 02/2021

Karlsruhe, Germany

Collaborated with an industry partner to address a domain-specific challenge for industry 4.0. Oversaw all technical aspects, including:

- **Hardware Development:** Designed and built an ESP-32-based IoT environmental sensor for automated daily data transmission to a remote server.
- **Web Application Development:** Developed a React-based web application to provide users with analysis and insights from the collected IoT data, hosted in an AWS environment.
- **Infrastructure Management:** Set up and managed cloud infrastructure for secure data storage and IoT device networking.

Technologies: AWS (VPC, EC2, DynamoDB, IAM, Route 53), React, MongoDB, PostgreSQL, Python, Typescript, NodeJs,

Projects

Cloud Community and Blog
DevLocus - (blog)
My blog and community platform for discussing the best practices in DevOps, MLOps and the Cloud.

An Evaluation Framework for Stateful Web App Testing
SWaTEval - (repo) (paper)
My contribution at Fraunhofer IOSB - a tool for web app state extraction to uncover vulnerabilities in web applications by analyzing complex state-dependent behaviors.

Coffee Type Classification from Vibrational Data
CoffAI - (link)
My contribution to the showcase project for Duesentrieb Lab. Credits go to esentri AG for the project.