BUGREEV ANTON

DevOps/MLOps Engineer

📍 Belgrade, Serbia
📞 +381621924980
✉ bugreev.anton@gmail.com
💬 https://t.me/vukor
💻 https://github.com/vukor
🔗 https://www.linkedin.com/in/vukor

ABOUT

DevOps/MLOps engineer with over 10 years of experience designing, automating, and scaling production infrastructure for machine learning platforms and mission-critical applications. Expert in Kubernetes (AWS EKS) workload migrations, cloud cost optimization, and performance tuning for large-scale ML inference. Proven track record of reducing deployment timelines, improving operational reliability, and enabling data science teams to move seamlessly from experimentation to production. Passionate about building efficient, reproducible ML workflows and integrating modern observability, CI/CD, and security practices to deliver scalable, compliant solutions.

TECHNOLOGIES AND TOOLS

- **Cloud Platforms:** AWS (EKS, RDS, S3, ElastiCache, EFS)
- **Containerization & Orchestration:** Docker, Kubernetes (AWS EKS, Rancher v1/v2)
- **Autoscaling**: Karpenter, KEDA
- **Infrastructure as Code:** Terraform, Ansible
- **CI/CD:** GitHub Actions, GitLab CI, Jenkins, Bamboo, ArgoCD, FluxCD, Helm, Kustomize
- **Monitoring & Observability:** Prometheus, Grafana, New Relic, CloudWatch
- **Logging:** Loki, Fluent Bit
- **Data & ML Pipelines:** Kubeflow Pipelines, Argo Workflows, Metaflow
- **Machine Learning & MLOps:** KServe, Knative, Langfuse, OpenWebUI + LiteLLM + Model Context Protocol (MCP), Label Studio, Weights & Biases
- **Development & Testing:** Python, FastAPI, JupyterHub, , K6 load testing
- **Security & Compliance:** Prisma Cloud, Snyk, Wallarm (HIPAA compliance)

WORK EXPERIENCE

Contract DevOps/MLOps Engineer

Belgrade, Serbia · 2022–Present

- Successfully migrated complex AI/ML infrastructure (Kubeflow Pipelines, JupyterHub, custom applications) from legacy environments to AWS EKS on the company's modern platform, reducing infrastructure costs and enhancing failover recovery capabilities.
- Integrated and tuned Karpenter autoscaling in production ML workloads, ensuring rapid burst capacity during high-volume inference and improving resource utilization efficiency.
- Developed Terraform IaC modules for automated provisioning of AI infrastructure in AWS, standardizing environments and reducing deployment setup time across teams.
- Migrated vector search databases from Vespa to OpenSearch; ran embedding model evaluations to select and deploy a solution that improved vector search accuracy by >15% for critical datasets.
- Automated Argo Workflows for retrieval-augmented generation (RAG) tasks, embedding generation, schema population, k6 load testing, and batch ingestion pipelines, resulting in higher reliability and faster data processing cycles.
- Deployed and managed Langfuse ML engineering platform for centralized prompt management; added LLM API request/response tracing to accelerate debugging and model iteration cycles.
- Contributed to OpenWebUI + LiteLLM + Model Context Protocol (MCP) component development, authoring system prompts and integrating AWS cost insights, GitHub, New Relic, and Prometheus, expanding operational visibility and intelligence for engineering teams.
- Wrote runbooks, incident playbooks, and operational guidelines, improving on-call response efficiency and MTTR for production incidents.

Quantumsoft - DevOps Engineer

Tomsk, Russia · 2016–2022

AI/ML project (2020-2022)

- Designed and operated multi-cluster Kubernetes environments (Rancher v2) supporting 50+ microservices, achieving 99.9% production uptime and ensuring high-availability for mission-critical workloads.
- Implemented scalable ML model serving with KServe, Knative, and KEDA; tuned autoscaling policies for performance efficiency.

- Developed Terraform modules to standardize provisioning of RDS, S3, Elasticache, EFS, and autoscaling node groups, reducing environment setup time from days to under 1 hour.
- Migrated CI/CD workflows from Jenkins to GitHub Actions, improving build speeds and deployment reliability.

Web Project (2016-2020)

- Automated end-to-end build, test, and deployment workflows using Ansible, Bamboo, and Rancher v1, reducing release time from 4+ hours to under 30 minutes and introducing blue/green deployment strategies to minimize downtime and release risk.
- Refactored monolithic application architecture into containerized microservices; moved workloads from Docker Compose to Rancher-managed orchestration, enhancing scalability and service isolation.
- Developed custom Ansible modules and playbooks for application builds and infrastructure provisioning, improving deployment consistency and reducing manual configuration errors.
- Integrated vulnerability scanning (Prisma, Snyk, Wallarm) to remediate 100+ critical/security CVEs, ensuring compliance with HIPAA PHI handling.

Online-media - Hosting System Administrator

Tomsk, Russia · 2012–2016

- Managed Linux/FreeBSD servers
- Automated hosting setup with Ansible and Docker for local dev environments

WebMedia - Hosting System Administrator (part-time)

Tomsk, Russia · 2009–2016

- Maintained Linux/FreeBSD systems
- Supported hosting and web/mail/vpn services

TOMTEL - Hosting System Administrator

Tomsk, Russia · 2006–2012

- Maintained Linux servers
- Supported hosting services and customer issues

EDUCATION

- Tomsk, Russia · 2001–2006
  - Tomsk University of Control Systems and Radioelectronics
  - Degree in Automated Control Systems, Computer-Aided Design

CERTIFICATIONS

- [LPIC-1 / LPIC-2](#)
- [RHCE](#)

LANGUAGES

Russian — Native
English — B2 (Upper-Intermediate)
Serbian - A2 (Elementary)