



We are hiring!

Senior MLOps Engineer (m/w/x)

We are looking for a **Senior MLOps Engineer** (m/f/x) to join our growing Data Science team. In this role, you will be responsible for maintaining and developing our Machine Learning Platform (MLP), optimizing machine learning models, and ensuring efficient deployment processes. You will work closely with cross-functional teams to drive innovation and improve the scalability of our ML solutions.



Key Responsibilities:

- Technical Leadership: Guide architectural decisions, set engineering standards, and mentor team members through daily collaboration
- Lead development of AIOps Capabilities— specifically capabilities for fine-tuning and serving of open-weights LLMs, model optimization and agentic AI
- Develop, deploy, and optimize ML models and algorithms using Google Cloud Platform (GCP) services such as Vertex AI, BigQuery, Cloud Storage, and Pub/Sub.
- Manage and resolve infrastructure-related issues, including failures in GitHub Actions, Docker builds, and disruptions in Vertex AI or BigQuery.
- Support incident management by providing quick and effective interventions in production issues.
- Optimize and restructure codebases to improve maintainability and efficiency.
- Oversee and manage Kafka topics used by our APIs.
- Provide and maintain VSCode workstations in the Google Cloud environment.
- Assist with deployments, particularly in structural changes such as schema adjustments.



Required Qualifications & Skills:

- Several years of experience with Google Cloud Platform (GCP), especially with Vertex AI, BigQuery, Cloud Storage, Pub/Sub, Cloud Functions, and Cloud Run or experience with similar cloud platforms
- Strong experience in Kafka.
- Expertise in Infrastructure-as-Code tools like Terraform or Deployment Manager.
- Deep understanding of the end-to-end ML lifecycle (MLOps) and model deployment principles.
- Proficiency in Python, especially in web frameworks (FastAPI, Flask, etc.) and familiarity with Go (beneficial).
- Hands-on experience with packaging, deploying, and serving models built in frameworks like PyTorch, TensorFlow, JAX or scikit-learn.
- Hands-on experience with LLMOps, including fine-tuning techniques and serving frameworks (e.g., vLLM, TGI, Hugging Face) for open-weight models.
- Hands-on experience with CI/CD pipelines like Jenkins, Tekton or OpenShift
- Familiarity with containerization and orchestration using Docker and Kubernetes.
- Familiarity with following observability tools: Grafana and Prometheus
- Proactive technical leader with strong problem solving skills
- Excellent communication skills and the ability to work in a team-oriented environment.

Apply now by sending us your CV to office@starbyte.com

We welcome applications from all individuals, regardless of origin, disability, gender, or orientation.

We are looking forward to receiving your application!