Aleksandar Babic

- **Location:** Serbia ● **Phone number:** +381643214300
- **Email address:** aleksandar@cloudhat.dev ● **Web:** https://ababic.cloud
- **Github:** aleksandar-babic ● **LinkedIn:** /in/aleksandar-babic

Profile

A seasoned Cloud Architect with 10+ years of experience designing and delivering secure, complex, high-performance systems across a wide range of industries, including FinTech, Automotive, Healthcare, Media, eCommerce, Web3, and AI. Proven track record in leading cloud transformation, platform engineering, and ML infrastructure initiatives for both high-growth startups and global enterprises. Expert in distributed systems, container orchestration, and automation, with deep knowledge spanning AWS, GCP, Kubernetes, and modern ML workflows. Recognized for sound technical judgment, execution under pressure, and building reliable systems that meet business-critical demands.

Experience

09/2024 – present

Principal Consultant
MyOps by Yael Group

- Recognized as the go-to architect for the company's most critical cloud initiatives, driving design and delivery across a wide range of clients and sectors
- Design and review high-assurance cloud environments with a strong emphasis on resilience, compliance, and cost-efficiency
- Support and mentor Solution Architects and Cloud Engineers across projects, providing deep technical guidance across infrastructure, DevOps, MLOps, and platform architecture
- Act as a trusted advisor to clients, often stepping in during critical phases of delivery to solve technical roadblocks and ensure systems meet production-grade standards
- Work hands-on with AWS, GCP, Kubernetes, and other modern technologies to deliver infrastructure and ML platforms that are stable, scalable, and easy to operate
- Own technical presales engagements, translating client goals into actionable architectures and presenting solutions that balance security, scalability, and performance

Industries: Artificial Intelligence (AI), SaaS, Startups, Fintech, Small & Medium-sized Enterprises (SME), Enterprises, Health Care, Automotive, Cutting-Edge
Technologies & Skills: AWS, GCP, Infrastructure as a Code, CI/CD Pipelines, Machine Learning, Deep Learning, FinOps, Kubernetes, GitOps, MLOps, Observability, Event Driven Architectures, Service Meshes

2021 – 09/2024

Solutions Architect
OpsGuru

- Helped SMBs and Enterprises shift to a cloud-native mindsets and processes.
- Worked with many Startups to accelerate the Cloud-native development processes.
- Designed highly available hybrid-cloud solutions for multiple enterprise clients.

Experience

- Sustained massive on-premises -> AWS migrations for Enterprises
- Supervised multiple teams of solutions architects and cloud engineers in implementing complex cloud-native green field and brownfield projects on AWS and GCP public cloud providers.
- Delivered complex, performance sensitive Kubernetes-based environments
- Designed and implemented many end-to-end MLOps solutions

Industries: Artificial Intelligence (AI), SaaS, Startups, Fintech, Small & Medium-sized Enterprises (SME), Enterprises, Health Care, Automotive, Cutting-Edge
Technologies & Skills: AWS, GCP, Infrastructure as a Code (Terraform/Terragrunt, CDK, Crossplane, Pulumi), CI/CD Pipelines, Machine Learning, Deep Learning, Kubernetes, GitOps, MLOps

2023 – 2024

DevOps Engineer
Lemongrass Consulting (US), Inc.

- Architected and implemented a SOC2 & HIPAA compliant GCP Landing Zone automation for a highly regulated production environment, used to serve 100s of Enterprise customers.
- Developed new features for the AWS Landing Zone automation.
- Conducted code reviews for team members on multiple platform components.

Industries: Business Services, Software Service Provider, SAP
Technologies & Skills: Python, Ansible, Terraform, Linux, Configuration Management, Terraform, Azure DevOps, Kubernetes

2022 – 2022

Cloud Architecture Consultant
Globaldatanet

- Architected a fully automated machine learning operations (MLOps) platform utilizing AWS cloud-native services to deliver models into production.
- Supervised a team of four cloud engineers in implementing the MLOps platform while following the security, quality, and performance best practices.
- Communicated directly with the client throughout the whole process.

Industries: Artificial Intelligence (AI), Pharmaceuticals, Software as a Service (SaaS)
Technologies & Skills: Machine Learning Operations (MLOps), MLflow, Kubernetes, TensorFlow, PyTorch, Kubeflow, Apache Airflow, Serverless, AWS Sagemaker, AWS Step Functions

01/2021 – 01/2022

AWS Cloud Engineer
Globaldatanet

• Architected and developed over 20 projects for a broad scale of clients varying from small startups to enterprises, with a focus on cloud-native, cost-effective solutions.
• Helped establish organization-wide code quality standards across multiple technology stacks.
• Implemented Cloud architectures with Infrastructure as a Code (CDK, Terraform, Terragrunt).
• Developed Python-based solutions for various projects.
• Regularly organized knowledge-sharing sessions focused on improving the quality of the utilization of cloud-native technologies.

Industries: SaaS, Startups, Fintech, Small & Medium-sized Enterprises (SME),

Experience

Enterprises, Health Care, Automotive
Technologies & Skills: AWS, Python, Terraform, CDK (Typescript), CI/CD, ML, Kubernetes

08/2019 – 01/2021

Senior AWS Cloud Engineer
Aegon Global Technology Services

- Architected highly available, scalable, enterprise-grade infrastructures in the cloud.
- Built a self-service portal for AWS account provisioning within the organization.
- Developed cloud-native, event-driven serverless microservices on AWS.
- Implemented CI/CD pipelines within various modules and services, helping to automate deployment processes and reduce time to market for new features.
- Applied Infrastructure as Code practices using Terraform and AWS CloudFormation.
- Managed AWS Organization with over 500 accounts, ensuring efficient governance and resource management across the entire organization.

Industries: Enterprises, Insurance
Technologies & Skills: AWS, CI/CD, Terraform, CloudFormation, Python, Kubernetes, AWS Step Functions, AWS ECS

01/2018 – 08/2019

Lead DevOps Engineer
Vivify Ideas

- Designed highly available, auto-scaling infrastructure for large-scale services on AWS.
- Managed cloud infrastructure using Infrastructure as Code (Terraform, CloudFormation) and Ansible for configuration management.
- Built containerized environments using Docker, orchestrated with Kubernetes and Docker Swarm.
- Developed automation scripts in Bash, Python, and Node.js.
- Led an agile team of 80 developers and established CI/CD pipelines for over 40 projects across multiple environments.

Industries: Startups, Small & Medium-sized Enterprises (SME), Fintech, Insurance, Health Care, Automotive, Cutting-Edge
Technologies & Skills: AWS, Python, Node.js, Bash, Terraform, CDK (Typescript), AWS, CI/CD, Kubernetes

01/2017 – 01/2018

Software Developer
Vivify Ideas

- Developed REST APIs using technologies like Laravel, Lumen, and ExpressJS, which enhanced the functionality of our applications and improved user experience.
- Built single-page and server-side rendered applications with JavaScript frameworks such as VueJS, Angular, and NuxtJS, streamlining the development process and boosting performance.
- Designed software architecture for various products, ensuring scalability and maintainability.
- Conducted thorough code reviews for all team members, helping to maintain high code quality and fostering a culture of continuous improvement within the team.

Industries: SaaS, Startups
Technologies & Skills: AWS, Javascript, Typescript, Node.js, PHP, Laravel, Lumen, ExpressJS, VueJS, Angular, NuxtJS, CI/CD

Experience

04/2016 – 06/2018

Linux System Administrator
PrecisionPros.com Network

- Managed and monitored OpenVZ containers, VPS, and bare-metal servers for hosting customers.
- Implemented automated OpenVZ container deployment and provisioning system.
- Automated routine administrative tasks using Ansible, Python, and Bash.
- Developed a distributed backup solution for a fleet of 140+ instances.
- Managed the production environments for various large-scale E-commerce platforms
- Resolved critical production system issues.

Industries: E-commerce, SMBs
Technologies & Skills: OpenVZ, Ansible, Python, Bash, Monitoring, PHP

Projects

Lossless Log Compression Platform

First-hire in a YC-backed startup, helped design and build the real-time, intelligent traffic analysis and lossless log compression platform that optimizes the log volume without dropping data, focused on the Datadog platform.

Was responsible for DSL Engine, Ingestion Pipelines, and optimization services. The platform was deployed in AWS EKS, utilizing Karpenter and Keda for robust and efficient auto-scaling.

Technologies & Skills: Python, Golang, Typescript, CDK, AWS, EKS, Karpenter, Keda, Nats

AWS Disaster Recovery Environment

Led the process of designing and implementing the SOC2, HIPAA, GDPR-compliant Disaster Recovery environment for the mission-critical, production workloads in AWS.

Was responsible for planning and delivering the multi-site active/active DR strategy, including 10s of microservices running in AWS EKS, EventBridge, PostgreSQL, Redis, S3, and EFS.

Additionally, prepared the DR Strategy Document, Service-level Runbooks, and DR Drill checklist.

Performed knowledge sharing sessions to ensure customer technical teams will adopt and perform regular DR drills.

Technologies & Skills: AWS, Terraform, Terragrunt, EKS, Kubernetes, Argo CD, GitHub Actions, EventBridge, PostgreSQL, Aurora, Redis, Elasticache, S3, EFS.

Kubernetes-Based Microservices Platform in GCP

Designed and implemented a production-grade, microservice-based platform in Google Cloud, hosting over 20 containerized services across multiple teams. Built on GKE with NATS as the messaging backbone for inter-service communication, the architecture leveraged KEDA for dynamic autoscaling based on workload metrics. Integrated managed services such as Cloud SQL, MemoryStore, and Cloud Storage for persistence, caching, and object

storage. CI/CD pipelines were implemented using GitHub Actions and Argo CD to ensure secure and consistent deployments. Instrumented services with OpenTelemetry and implemented observability with the LGTM (Loki, Grafana, Tempo, Mimir) stack for centralized logging, tracing, and metrics.

Technologies & Skills: GCP, GKE, Kubernetes, NATS, KEDA, Cloud SQL, MemoryStore, Cloud Storage, Terraform, Argo CD, GitHub Actions, OpenTelemetry, Loki, Grafana, Tempo, Mimir.

Real-Time Stable Diffusion Pipeline Optimization

Developed a platform that enables clients to deploy multiple diffusion-based pipelines in production, which significantly improved their operational efficiency. Utilized Nvidia Triton as a production-grade inference server, allowing clients to scale from a single concurrent request per GPU to eight, enhancing their processing capabilities. Built an automated system to convert existing pipelines into Triton-compatible formats, which improved both inference times and concurrency for better performance. Facilitated customer growth and acquisition of new enterprise clients by unblocking their scaling limitations, ensuring they could handle increased loads seamlessly.

Technologies & Skills: AWS, Kubernetes, Nvidia Triton, PyTorch, TensorRT, ONNX, Step Functions, Python, ArgoCD, Argo Workflows

Production-grade LLM Platform

Designed and developed a proprietary platform that enables the self-hosting of LLMs on Kubernetes, ensuring a highly available, auto-scalable, and self-healing environment, which enhances reliability and performance for users. The solution allowes customers to deploy Qwen 2.5 VL, Qwen 3, and various LLama LLMs without incurring token-based costs, helping to reduce financial barriers to access advanced language models. Achieved an impressive 80% cost reduction compared to previous expenses with external LLM providers, which significantly improved budget efficiency for the organization.

Technologies & Skills: AWS, Kubernetes, Terraform, vLLM, Keda, SQS, Cuda, Helm, ArgoCD, GitOps

Deep Learning Platform

Architected and led the process from discovery and design to implement a deep learning platform based on the gRPC Kubernetes-based microservices. Utilized state-of-the-art technologies on GCP and extremely compute-intensive resources with Cuda GPUs. The platform is entirely cloud-native, stateless, and managed through infrastructure as code following GitOps principles.

Technologies & Skills: Python, gRPC, Google Cloud Platform (GCP), Google Kubernetes Engine (GKE), GitHub Actions, TensorFlow, Kubernetes, Keda, Cuda, MLFlow, Vertex AI

MLOps Platform

Architected a fully automated MLOps platform utilizing AWS cloud-native services to deliver 30+ ML models from development into production seamlessly. Supervised a team of four cloud engineers in implementing the MLOps platform while following security, quality, and performance best practices.

Technologies & Skills: Amazon Web Services (AWS), Terraform, Python,

Projects

Amazon SageMaker, Apache Airflow, MLflow, Kubeflow

Cloud Factory

Developed a fully automated AWS account factory solution that provisions and maintains a best practices-driven, multi-account environment allowing businesses to rapidly start their cloud journey while ensuring that they follow all industry standards.

Technologies & Skills: Python, Golang, AWS Service Catalog, AWS Step Functions, AWS Lambda, AWS ECS, Terraform

Certifications

Solutions Architect Professional
Amazon Web Services

Professional Cloud Architect
Google Cloud

Solutions Architect Assocaite
Amazon Web Services

Developer Associate
Amazon Web Services

Associate Cloud Engineer
Google Cloud

Skills & Tech

Cloud	Cloud Security	AWS	GCP
FinOps	DevOps	DevSecOps	MLOps
GitOps	Cloud-Native Architectures	Event Driven Architectures	Cloud Governance
Cloud Cost Optimization	Zero Trust	Service Meshes	Microservices
Kubernetes	Kubernetes Operator Patterns	CI/CD	CNCF
Infrastructure as a Code	Terraform	CDK	Python
Node.js	Golang	PHP	Argo CD
FluxCD	Observability	OpenTelemetry	APM
Istio	Linkerd	Nginx	Ingress-nginx
Traefik	Cert-manager	External-DNS	Step Functions

Skills & Tech

Serverless	CDKTF	CDK8s	CloudFormation
AWS SAM	Pulumi	EKS	GKE
Helm	Containerd	Docker	Typescript
Javascript	Ruby	Argo Workflows	Argo Rollouts
Github Actions	Gitlab CI	Tekton	AWS CodePipeline
AWS CodeBuild	GCP Cloud Build	CircleCI	Jenkis
Amazon Sagemaker Ecosystem	GCP Vertex AI Ecosystem	MLFlow	vLLM
KubeFlow	TensorFlow	Hugging Face	Transformers
PyTorch	Apache Airflow	Databricks	DVC
Amazon Aurora	Amazon RDS	GCP Cloud SQL	PostgreSQL
MySQL	MariaDB	MongoDB	MSSQL
Oracle	DynamoDB	GCP Datastore	GCP MemoryStore
AWS ElastiCache	Amazon DocumentDB	Timescale DB	Amazon Timestream
Elasticsearch	Neo4j	Cassandra	Athena
BigQuery	Kinesis	Redis	Valkey
InfluxDB	Prometheus	Grafana	Loki
Mimir	Grafana Alloy	RabbitMQ	Nats
Kafka	KubeMQ	SQS	Amazon SNS
Amazon MQ	GCP PubSub	gRPC	GraphQL
Rest	Amazon Cloudwatch	GCP Cloud Logging	DataDog
NewRelic	ELK Stack	LGTM Stack	Zabbix
Nagios	Bash	Powershell	... and many others

Education

2015 – 2019
Novi Sad, Serbia

Information Technology | Bachelor's Degree
University of Novi Sad