Ashis Sharma

Hyderabad, India 500039
+91 8759150942
ashissharma@outlook.com

Summary

DevOps and AI Infrastructure Engineer with 4+ years of experience deploying large-scale LLMs and building scalable CI/CD and MLOps pipelines. Proven success in optimizing cloud workloads, orchestrating high-performance GPU clusters, and automating secure DevSecOps pipelines in production environments.

Skills

- Cloud Platforms: Google Cloud, Azure
- Containers & Orchestration: Docker, Kubernetes, Slurm, K3s, Podman
- Programming Languages: Python, GoLang, JavaScript, Bash
- CI/CD & Infrastructure: GitLab CI/CD, GitHub Actions, Terraform
- Monitoring & Security: Prometheus, Grafana, EFK Stack, DevSecOps tools
- HPC & AI Tools: CUDA, TensorRT, NCCL, vLLM

Experience

Jukshio June 2020 to Current

DevOps Engineer

Hyderabad, India

- Fine-tuned and deployed ultra-large LLMs including **DeepSeek R1/V1-671B** and **LLaMA 3 405B** across distributed **HPC GPU clusters** with ~160 **NVIDIA A100s**, achieving scalable multi-node performance.
- Designed and implemented high-throughput **inference pipelines** for large models, optimizing latency and throughput in **GPU-based distributed systems**.
- Benchmarked and profiled model performance across **multi-node environments**, identifying bottlenecks and improving **training/inference efficiency** by over 20%.
- **Developed a scalable Kubernetes cluster deployment for Raspberry Pi devices**, enabling remote IoT updates and resilient edge computing environments.
- Led the **cloud migration** of critical workloads from **Microsoft Azure** to **Google Cloud Platform (GCP)**, optimizing resource allocation and reducing infrastructure overhead.
- Developed and integrated a **DevSecOps pipeline** for a high-security web application, incorporating static code analysis, DAST, dependency scanning, and infrastructure security testing.
- Created scalable **CI/CD pipelines** using **GitLab CI** and **Terraform**, reducing deployment cycles by 60% and ensuring consistent, reproducible environments.
- Orchestrated containerized workloads using **Kubernetes** and **Docker**, maintaining high availability and seamless scalability for production services.
- Implemented robust **monitoring and logging solutions** using **Prometheus**, **Grafana**, **Loki**, and the **EFK Stack**, enhancing system observability and reliability.
- **Built and deployed a Kubeflow-based pipeline** to streamline the training of deep learning models with **TensorFlow** and **PyTorch**, including installation and configuration of **high-performance GPU drivers** for optimal compute efficiency.
- Developed **SDKs** for enterprise clients such as **Jio** and **HDFC Bank**, providing seamless integration and tailored technical support.
- Created an internal tool using **React** and **Node.js**, improving operational efficiency and cross-team collaboration.
- Maintained and optimized the company's public-facing website using modern **full-stack web technologies**, boosting performance and usability.

Wipro August 2019 to March 2020
Test Engineer
Chennai, India

- Design, develop, and maintain automated test scripts using **Selenium WebDriver** in Java for web-based applications.
- Execute test cases on mainframe systems, and extract logs and reports generated transaction programs.
- Develop **Java-based utilities** to parse data and generate Excel reports

Education

SRM University Sikkim May 2019
Bachelor of Science: **Information Technology**
Gangtok

Websites, Portfolios, Profiles

- linkedin.com/in/ashissharma/
- github.com/ashissharma97
- https://medium.com/@ashisrm