Vinay Gudi

Site Reliability Engineer II

Experienced DevOps Engineer with a demonstrated history of working in the information technology and services industry. Skilled in DevOps, Configuration Management, Linux, Amazon Web Services (AWS), and Shell Scripting. Professional goal is to work in a globally competitive environment on challenging assignments that shall yield the twin benefits of the job satisfaction and a steady paced professional growth.

✉ gudivinaykumar98@gmail.com ☎ 8146399973 HYDERABAD, India

EDUCATION

B.Tech (Computer Science and Engineering)

Lovely Professional University

08/2016 - 08/2020 *Phagwara, Punjab, India*

CERTIFICATES

Site Reliability Engineering: Measuring and Managing Reliability(Coursera)

HashiCorp Certified: Terraform Associate (002)

MongoDB Certified DBA Associate (C100DBA)

AWS Certified Solutions Architect - Associate (SAA)

Introduction to GitOps(LFS169)

Big Data Engineering with Hadoop and Spark(cloudxlab)

WORK EXPERIENCE

Site Reliability Engineer II

Chegg Inc.

Achievements/Tasks

- Engaged with our product teams to understand requirements, design, and implement resilient and scalable infrastructure solutions.
- Operated, monitored, and triaged all aspects of our production and non-production environments.
- Evaluated and integrated new technologies to improve system reliability, security, and performance.
- Developed and implemented automation to provision, configure, deploy, and monitor various chegg services.
- Participated in an on-call rotation providing hands-on technical expertise during service-impacting events. Contributed to capacity planning, scale testing, and disaster recovery exercises Approach operational problems with a software engineering mindset.
- Building and operating container orchestrating systems like Kubernetes or EKS.
- Performed detailed RCA of issues with proper documentation.
- Troubleshooted issues across the entire stack – infrastructure, software, application and network.
- Worked with the teams to identify ways to increase MTBF and lower MTTR for the environment.
- Maintaining environment monitoring systems to provide the best visibility into the state of the deployed products/solutions.

WORK EXPERIENCE

Senior DevOps Engineer

Inviz AI

06/2020 - 10/2022

Achievements/Tasks

- Worked on multiple client projects as DevOps engineer from development to production.
- Managed a team of three DevOps engineers and Implemented and maintained Continuous Integration/Continuous Delivery systems.
- Represented DevOps team in client meetings and demos. Built, maintained, and scaled infrastructure for production, QA, and dev environments.
- Deployed and maintained Highly Available Large Elasticsearch cluster in kubernetes from dev to production loads using elastic operator and monitoring elastic cluster.
- Tested and maintained event driven machine learning model serving using knative eventing and knative serving.
- Performed monitoring and logging using various tools like prometheus, grafana and datadog

DevOps Engineer

Inviz AI

08/2019 - 10/2022

Achievements/Tasks

- Worked on setting up infrastructure on GCP and AWS using terraform.
- Developed and deployed various micro-services on kubernetes and docker using helm for version control.
- Designed and deployed basic reliable and highly available architecture on AWS.
- Shifted monitoring and logging to Datadog from prometheus stack.
- Developed a complete cloud native storage solution in kubernetes with rook.

PROJECTS

LoopR (Inviz AI)

- Designed architecture and wrote the terraform scripts for provisioning of multiple environments.
- Developed and maintained CI/CD pipelines using gitlab.
- Learnt and used various open sources cncf projects like velero,minio,harbor and rook.
- Used istio as service mesh to manage service to service authentication and networking.
- Setted up autoscaling and metrics required for the production environment.

ArboHub (Panasonic)

- Maintained production servers running on multiple EC2 nodes.
- Worked on mongoDB index and queries performance optimization.
- Reduced costs by 20% using various cost cutting methods.
- Migrated applications from EC2 fleet to kubernetes clusters without any major downtime.
- Deployed, scaled and maintained various opensource cncf projects in multiple environments - istio, prometheus, harbor and velero.

Search platform (Tata Cliq)

- Designed initial architecture and wrote the terraform modules for provisioning of multiple environments.
- Developed CI/CD pipelines spanning multiple AWS accounts using AWS code pipeline.
- Designed Highly available cross accounts AWS architecture working closely with Amazon SME.
- Deployed and maintained Highly Available Large Elasticsearch in kubernetes from dev to production loads using elastic operator and monitoring elastic cluster.
- Designed and deployed machine to machine authentication and authorization using AWS cognito and api gateway.
- Deployed kubeflow for training and serving ML models in production with kf-serving and knative.
- Deployed on-premise kubernetes cluster of 5 nodes using microk8s and rook-ceph for storage pools.
- Worked very closely with TATA's devops team and senior managers for architectural and performance approvals of various services.
- Reduced existing AWS costs by ~30% by using various cost cutting methodologies.

PROJECTS

API Infra Template (Chegg)

- Built a Massively Scalable API Gateway handling millions of requests/day with Istio + Envoy + Kong.
- Designed DDoS-resistant architecture with Cilium (eBPF), blocking 10M+ RPS attack vectors.
- Used open-sourced high-performance, zero-allocation logging libraries , standardizing log formats across applications and improving log readability.
- Enforced Zero Trust security using Keycloak, Istio mTLS, and OAuth2.0 token validation.
- Integrated Rate Limiting (Istio), WAF (Kong ModSecurity), Bot Detection, and Geo-IP Blocking.
- Reduced API latency by ~80% using Envoy caching + Redis for frequently accessed responses.
- Built Event-Driven Processing with Kafka + KEDA, auto-scaling based on request load.
- Achieved 99.99% uptime using Multi-Cloud HA deployments + Velero Disaster Recovery.
- Monitored Kubernetes runtime security with Falco, detecting unauthorized container actions.

Production Incident Automation and Optimization (Chegg)

- Designed and implemented an automated incident response framework to address the top 50 repetitive production incidents, improving incident resolution efficiency and reducing on-call engineer workload.
- Analyzed three years of production incidents to identify recurring issues and developed targeted automation jobs for efficient incident resolution.
- Integrated the automation system with alerting tools to trigger incident-specific jobs that gather logs, metrics, and service dependency data (upstream and downstream) in real-time.
- Implemented anomaly detection algorithms to identify unusual patterns in logs and metrics across affected services during incident response.
- Built dynamic dashboards using tools like Grafana or Kibana to visualize key incident-related data, enabling quicker root cause analysis and sharing with stakeholders.
- Automated the creation of Jira tickets enriched with detailed logs, metrics, anomaly reports, and dashboard links for seamless post-mortem analysis.
- Developed email notification systems to alert on-call engineers with incident details and dashboard links, enabling faster response times.
- Pogrammatically set up Slack channels with predefined participants (engineers, team leads, and stakeholders) to streamline real-time communication during incidents.
- Reduced mean time to resolution (MTTR) by integrating automated workflows across monitoring tools, incident management platforms, and collaboration tools.
- Ensured the framework is modular and scalable to accommodate new incident types and integrate with evolving monitoring and alerting tools.

Event-Driven MLOps Setup (Chegg)

- Designed a highly scalable event-driven architecture powered by Kafka to handle real-time streaming data for model training and prediction workflows. Integrated these pipelines with Kubeflow and KFServing for real-time inferencing with advanced autoscaling capabilities.
- Orchestrated multi-cluster Kubernetes deployments using FluxCD, enabling declarative configurations, continuous deployment, and instant rollback capabilities, while managing complex dependencies between microservices.
- Deployed Istio to enforce fine-grained traffic control, service-to-service authentication, and rate-limiting for critical workloads. Integrated Jaeger to perform distributed tracing and ensure performance optimization of AI microservices.
- Engineered a distributed object storage system using Rook-Ceph, enabling seamless data sharing across clusters. Optimized performance for AI model checkpoints and high-velocity streaming data.
- Implemented Velero to manage scheduled and ad-hoc backups across production environments, ensuring rapid restoration during incidents without compromising data consistency.
- Integrated Auth0 to provide enterprise-grade authentication and authorization, including SSO, OAuth2, MFA, and RBAC policies, securing access to ML pipelines, APIs, and dashboards.
- Built real-time monitoring dashboards with Prometheus, Grafana, and NewRelic to measure SLAs, SLOs, and error budgets. Ensured proactive incident response by implementing AI-based anomaly detection using custom Prometheus rules.
- Automated the entire lifecycle of ML models, including training, versioning, deployment, monitoring, and decommissioning. Set up model drift detection and feedback loops for continuous improvement.
- Designed and tested disaster recovery procedures for MongoDB and MySQL on Kubernetes, ensuring that database clusters could be restored efficiently in case of failure using point-in-time restores from Velero snapshots and custom scripts.Developed custom Kubernetes Operators to automate common database administrative tasks for MongoDB and MySQL, including automatic scaling, backup management, and cluster health checks, improving operational efficiency and reducing manual intervention.
- Set up multi-cluster synchronization for MongoDB and MySQL using replication and sync tools.Conducted performance tuning on MongoDB and MySQL databases running on Kubernetes

SKILLS

GIT · Kubernetes · Linux · Docker · AWS · istio · Rook · minio · velero · python · pyspark · bash · prometheus · Newrelic · nginx · ELK · API Gateway · KAFKA