Ujwal Kandi

(737) 326-8598 Austin, Texas ujwal.kandi@utexas.edu linkedin.com/in/ujwalkandi www.ujwalkandi.in

Professional Summary

Site Reliability & Observability Engineer with 3+ years of proven expertise and an M.S. in Information Technology & Management from UT Austin. Excel at transforming operational challenges into scalable solutions that drive business continuity and cost optimization across production systems. Architect monitoring and incident response frameworks reducing MTTR by 40%+, preventing 60%+ of incidents pre-impact, and maintaining 99.9% SLA compliance. Combine infrastructure automation with AI/ML for intelligent edge diagnostics, delivering measurable ROI through reliability engineering. AWS/Azure proficient across 16+ production environments with demonstrated impact in disaster recovery, distributed systems optimization, and operational efficiency improvements by 35%+.

Experience

Apple Inc. (Contract via Welo Data) **Jan 2025 – Present**
Metadata Engineer *Austin, TX*

- Optimized mission-critical metadata ingestion pipelines ensuring 99.9%+ uptime and <2s latency for Apple TV's global content catalog using automated validation, error handling, and rollback procedures across **XML** workflows
- Established monitoring and alerting for metadata pipeline health to detect and triage data quality issues, reducing feed failures by 30% and improving platform stability
- Implemented structured debugging and anomaly detection frameworks to identify root causes of batch job failures, reducing **mean time to resolution** (MTTR) by 25% through systematic RCA and runbook development
- Designed disaster recovery procedures for critical metadata systems, ensuring continuity and compliance with Apple's SLA requirements

Sports Excitement **Aug 2024 – Jan 2025**
DataOps Engineer *Austin, TX*

- Built and maintained production ETL pipelines using Apache Airflow (DAG orchestration, monitoring, and failure recovery), supporting 50+ daily batch jobs with 99.9% reliability
- Optimized Azure Data Lake Storage architecture, reducing query latency by 35% and improving downstream ML analytics throughput
- Automated incident notifications via Airflow trigger jobs and alerting workflows, achieving <5min response times within SLA
- Implemented data quality checks and load-testing frameworks, reducing post-deployment failures by 30%

Dover Fueling Solutions **Jun 2024 – Aug 2024**
Jr Site Reliability Engineer *Austin, TX*

- Enhanced operational observability by implementing 3-pillar monitoring strategy: configured Grafana dashboards, Azure APIM log analytics integration, and Kafka metrics exporters; reduced anomaly detection time from 15min to 3min and MTTR by 40%
- Designed and deployed agentic AI system for edge device diagnostics on resource-constrained hardware (SQLite + Ollama), enabling field technicians to resolve issues autonomously, reducing support tickets by 35% and improving MTTR for critical faults to <10min
- Utilized LangChain to prototype multi-agent configurations and employed LangSmith to evaluate agent performance, response accuracy, and token efficiency in edge inference scenarios while maintaining diagnostic accuracy under stringent 2GB memory constraints
- Implemented alert tuning and intelligent escalation workflows to reduce alert fatigue by 45% while maintaining 99.9%+ critical incident detection rate; documented best practices in runbooks for on-call rotations

Stronghold Investment Management **Jan 2024 – May 2024**
AI Engineer - Capstone Project *Austin, TX*

- Designed and implemented a custom-built AI agent over Azure AI Studio, integrating OCR and LLM (Mixtral) models to enhance ownership verification processes and automate complex document workflows
- Reduced document processing costs by 20% by implementing an automated document intelligence system for document analysis
- Conducted model response evaluations across multiple Azure-hosted models using custom test datasets and task-specific tuning to benchmark performance (accuracy, coherence, latency) across query types
- Structured prompt design evaluations within Azure AI Studio to refine templates and optimize response quality with prompt tuning for document analysis workflows

Epsilon **Apr 2022 – Jul 2023**
Production Support Engineer *Bengaluru, India*

- Orchestrated production incident response lifecycle (detection → triage → resolution → RCA → prevention) across 16+ client systems, achieving 99.9% SLA compliance and reducing MTTR by 40% through proactive monitoring and automated remediation
- Designed and maintained comprehensive monitoring infrastructure: integrated **CloudWatch**, Kibana, and Dynatrace; established intelligent baselines, thresholds, and anomaly detection algorithms preventing 60% of potential incidents before customer impact
- Authored 20+ runbooks, SOPs, and postmortem documentation implementing blameless RCA culture; broke repeat incident cycles by 50% through systematic root cause prevention and knowledge transfer
- Implemented Infrastructure as Code disaster recovery framework: automated failover scripts for ETL pipelines using CloudFormation, reducing recovery time objective (RTO) from 2hrs to 30min and maintaining 99.9% uptime during critical incidents
- Optimized Matillion ETL job orchestration (SLA schedules, retry logic, maintenance windows), improving batch completion efficiency by 35% and reducing operational overhead by 40%
- Resolved critical data load failures within Epsilon's data mart by optimizing complex batch processes, improving system performance by 89% and ensuring consistent data availability across 99.9% of daily loads

Epsilon **Aug 2021 – Apr 2022**
Data Programmer I *Bengaluru, India*

- Engineered and optimized ETL pipelines for 16+ global clients (Unilever, Marriott, Bank of America, Disney, National Geographic), improving data load reliability and reducing batch failure frequency by 45%
- Designed test pipelines and production-like staging environments to simulate high-volume data flows, reducing post-deployment issues by 35% and improving validation accuracy through systematic performance testing
- Automated 50+ ETL batch jobs and implemented CloudWatch/Dynatrace/Kibana monitoring, enabling early anomaly detection and accelerating root cause identification by 30%
- Resolved critical data ingestion and quality issues across AWS S3 and Redshift environments, achieving 99.9% data availability for mission-critical analytics platforms

Isthriwala **Jun 2020 – Jul 2021**
Data Management Analyst *Vijayawada, India*

- Optimized database operations for 35,000+ individuals and 50+ B2B clients using SQL, improving query performance
- Developed KPI dashboards in Tableau for real-time business metrics, improving decision-making velocity and enabling 15% improvement in monthly performance targets

Technical Skills

Languages & Scripting: Python, Bash, SQL, YAML, XML, JavaScript, R

Observability & Monitoring: Grafana, Prometheus, CloudWatch, Kibana, Dynatrace, Azure Insights, ELK Stack

Infrastructure & Deployment: Kubernetes, Docker, Terraform, CloudFormation, Ansible

Data Pipelines & CI/CD: Apache Airflow, Kafka, Jenkins, GitLab CI/CD

Cloud Platforms: AWS (EC2, S3, Lambda, RDS, Redshift, CloudWatch), Azure (ADLS, Log Analytics, AI Studio, APIM)

AI/ML & LLM: LangChain, Ollama, LangSmith, ChromaDB

Tools & Platforms: Git, GitHub, Jira, ServiceNow, Apache Spark, PySpark, Tableau, Power BI

Education

The University of Texas at Austin **2023 – 2024**
M.S. in Information Technology & Management *Austin, TX*
McCombs School of Business
Coursework: Distributed Systems, Data Management, Unstructured Data Analytics, Data Governance & Responsible AI

Projects

A.I.D.E. – AI-Powered Observability & RCA Engine

- Engineered end-to-end observability solution combining metrics analysis, log anomaly detection, and AI-driven root cause identification for API infrastructure troubleshooting using multi-agent system (LangChain) with tool-calling architecture enabling autonomous log parsing, pattern recognition, and hypothesis testing
- Optimized LLM inference for edge deployment: integrated Ollama + Groq achieving sub-second diagnostic latency under resource constraints (<2GB memory), enabling on-site diagnostics without cloud dependency
- Deployed with RAG framework (ChromaDB) for contextual knowledge integration, ensuring diagnostic accuracy matches production monitoring systems
- Built during ATX Llama Hackathon (Meta), utilizing Llama Stack for on-device intelligence and low-latency inference

Yelp Reviews Recommender System

- Engineered recommender system using custom web scraper and NLTK sentiment analysis for data-driven business intelligence
- Delivered actionable insights to consulting team, optimizing menu and service offerings for 3 restaurants based on customer sentiment patterns and recommendation algorithms

Leadership & Achievements

Seeed Studio Ranger Program **Oct 2024 – Present**
AIoT Developer Program *Austin, TX*

- Designed and delivered hands-on AI/IoT technical training workshops reaching 100+ university students, building practical skills in artificial intelligence and edge computing systems.

Dover Fueling Solutions Hackathon'24 – First Place Winner **Feb 2024**
Annual Hackathon *Austin, TX*

- Won company hackathon designing conversational AI system with edge computing for fuel station operations; presented solution at DFS User Conference 2024 to executive leadership and industry stakeholders.

MSITM Campus Ambassador – McCombs School of Business **Aug 2023 – May 2024**
Graduate Program Ambassador *UT Austin*

- Guided prospective students through the application process, increasing applications from underrepresented backgrounds by 60%.

University Innovation Fellows **Jan 2018 – Dec 2020**
International Fellowship Program *Stanford d.school*

- Conducted 15+ workshops on Design Thinking and problem-solving reaching 1000+ students across multiple universities; earned scholarship to attend Silicon Valley immersive learning experiences at Stanford, Google, and Microsoft.