Soumendra Kumar Sahoo

AI Observability Architect | AI Solutions Sustain Lead

@ soumendrak@hotmail.com ⌘ www.soumendrak.com ⌨ @soumendrak in soumendrak ⊚ Hyderabad, India

AI Observability & ML Engineering Leader with 13 years building enterprise-scale observability stacks, back-end services, and production ML platforms. Proven expertise in model evaluation, monitoring, and alerting across traditional and GenAI (8 yrs), plus 3 yrs hands-on in deploying and scaling live models. Led cross-functional teams of 20+ engineers (dev, QA, DevOps) to deliver end-to-end solutions.

Experience

AI Observability Architect & AI Solutions Sustain Lead

PepsiCo - PepVigil

📅 2024 Dec. — Present   ⊚ Hyderabad, India

- Architected & launched PepsiCo's unified AI Observability platform, spanning LLM/Agents, computer-vision models, and traditional ML across global business units. The platform is handling 1 billion+ traces per month.
- Operationalized production Agentic solutions by adding deep instrumentation, real-time monitoring, alerting, and performance tracking —sharply reducing incident MTTR and boosting model trust.
- Directed a 10-plus cross-functional team of developers and QA engineers to deliver high-availability, large-scale observability services that support mission-critical AI solutions.

Lead Systems Engineer

Freshworks - Freddy AI

📅 2021 Sep. — 2024 Nov.   ⊚ Bengaluru, India

- Designed end-to-end MLOps platform for Model Monitoring, Observability, Logging, Evaluation, Versioning, Usage, and Adoption tracking.
- Developed and designed dashboards for monitoring the models availability, infra usage, latency, efficacy, and throughput.
- Scaled up the throughput of traditional ML services from 50 to 1000 requests per sec.
- Debugged end-to-end applications to hunt down bottlenecks and memory leaks, reducing the p99 latency from 15 minutes to under a second.
- Optimized large table (> 2TBs) MySQL queries, scaled up Database bottlenecks using sharding/partitions, and reduced costs by removing old records.
- Other than the unit and sanity tests, performance tested the ML application by mimicking the expected production load in the staging environment.
- Made extensive documentation and design on all the proposed and implemented changes.
- Analyzed every P0/1/2 outage received on existing applications and made permanent fixes to the root cause, resulting in zero P0 in the last year and only two P1 issues in the application.
- Designing the next MLOps infrastructure for the entire org from Data Engineering to Model Deployment and Model Retraining pipeline.

AIOps Platform Lead

IBM - Multi Cloud Managed Platform

📅 2018 Dec. — 2021 Aug.   ⊚ Bengaluru, India

- Implemented a consumption-based pricing model for clients, which enabled a pay-as-you-go model. Also, I implemented multiple pricing tiers, allowing the clients to use different versions per their use cases.
- Architected and implemented end-to-end user authentication and authorization in the ELK stack.

Education

BITS, Pilani, India

2021 - 2023   Correspondance

M.Tech in Data Science and Engineering
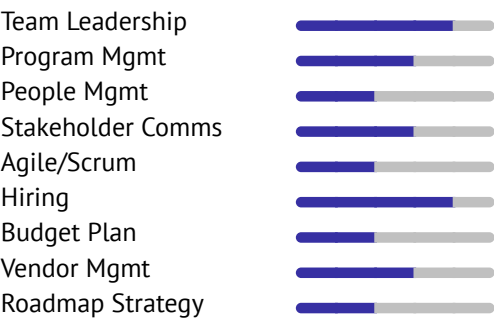
ITER, S'O'A University, Bhubaneswar, India

2008 - 2012   Bhubaneswar, India

B.Tech in Electrical Engineering

Technical Expertise

Python	
Go	
Docker	
SQL	
Databricks	
CI/CD	

Managerial Expertise

Team Leadership	
Program Mgmt	
People Mgmt	
Stakeholder Comms	
Agile/Scrum	
Hiring	
Budget Plan	
Vendor Mgmt	
Roadmap Strategy	

Programming Languages

• Python (Flask, Django, Litestar, FastAPI) • Go (Fiber, Gin) • Javascript/Typescript (ReactJS) • Shell script

AI/ML Tools

• Arize • Databricks • Spark • OpenAI API • MLFlow • LangChain • Kubernetes • Docker • Git • Jenkins • GitHub Actions • Kibana • Grafana

Databases

• MySQL • PostgreSQL • AWS ElastiCache • MongoDB • Redis • Elasticsearch • IBM DB2

Cloud Experience

• AWS • Azure • Digital Ocean • IBM

- Led two developers and one QA engineer and gained domain knowledge on AIOps and its use cases.

Technical Lead

Wipro - Cognitive Search

🗓 2016 Oct. – 2018 Dec.   📍 Bengaluru, India

- Deployed an enterprise level search engine from PoC to a scalable production level for 1000 concurrent users.
- Led a team of 10 developers and QA engineers to develop an enterprise-grade search engine from scratch.

Application Development Analyst

Accenture - Finance Transaction

🗓 2015 May. – 2016 Sep.   📍 Bengaluru, India

- Automated complex human screen entries using VB Script and saved worth $40,000+ per year time of maintainers.
- Gathered and broke down the requirements into manageable problems to design solutions while working with one of the top three banks in the world and gained immense domain knowledge on money transfer.

Systems Engineer

Tata Consultancy Services - Securities Statements

🗓 2012 Jun. – 2015 May.   📍 Mumbai, India

- Worked with one of the top stock brokerage firms in the USA. I learned about mutual funds and stocks.
- Organized and participated with the team to create the strategy for the enhancement and develop the optimum plan.
- Solved 50+ bugs and maintained old legacy finance applications.

Volunteering/Pro bono Consulting

AI Consultant

Stealth - AI Solutions

🗓 2023 Nov. – Present   📍 Remote

- Designed and developed multi-tenant chat assistance to reduce human efforts and increase the number of leads for a startup in the education domain, resulting in $10,000 per year in cost savings.
- Designed and consulted on a Voicebot to make outgoing sales calls and generate leads.
- Consulted on various tools like Email ID validator, Call recorder, etc.

CFP Lead and Core Team Member

PyCon India - Annual Conference

🗓 2023 Apr. – 2023 Oct   📍 Hyderabad, India

- Lead the activities for the CFP work group for the smooth conduct of the conference. Supervising Talks selection, Speakers, Workshops, and Posters proposals

Co-Founder and Core Team Member

Odisha AI - Annual Conference

🗓 2020 Jul. – Present   📍 Bengaluru, India

- Lead the logistics activities, took important decisions and managed crisis situations while organizing one of the longest 33hrs conference.

Other Tools

• Kafka/RabbitMQ • Nifi • Airflow • Nginx • Envoy Proxy • Caddy • Bruno/Postman

Documentation Tools

• Markdown • Typst • GitHub pages • Cloudflare Pages • Zola • Jekyll • Confluence • Jupyter Notebook • Excalidraw • MS Visio • D2 • Draw.io • DrawDB

Achievements/Certifications

AWS ML Specialty

- Got skills on AWS Machine learning products and certified.

Azure AI Fundamentals

- Got basic AI product skills on Azure Cloud and certified.

Social Impact Champion

- Outstanding contribution to the community by being core team members of OdishaAI, OdiaGenAI, Mozilla Common Voice and admin of Odia Wikipedia. Have 5+ million followers across social medias.

Languages

English
Hindi
Odia
German

Other Personal Skills

OpenOdia Python Package

- Maintainer of an open-source PyPi package, OpenOdia, with 40,000+ downloads.

Shabdarasa

- Developed Shabdarasa, a wordle game for the Odia language with 100+ DAU in ReactJS and Tailwind CSS.

OpenStreetMap

- Mapped over 3500 kilometers of roads in India and 75% of the buildings in Bhubaneswar.