Soumendra Kumar Sahoo

Lead MLOps Engineer

Email: soumendrak@hotmail.com
Current Address: Bengaluru, India
Phone Number: +91-9902437522
LinkedIn: /in/soumendrak
Website: soumendrak.com
Recommendation: Testimonials

JOB SKILLS

- Accomplished professional with extensive experience in scaling an ML service, back-end software development, and leading teams to develop end-to-end solutions in Banking & Finance, and Education domains.
- 3+ years of experience in deploying, operationalizing, maintaining and scaling **ML models in production**.
- 2.5 years of experience in an **end-to-end enterprise level search engine development** in Wipro**.** 3+ months of experience on RAG based search.
- 8+ years of experience in **both traditional and Generative AI models**.
- 12+ years of experience in multiple phases of Software Development Life Cycle including coding standards, code reviews, source control management, build processes, testing, and operations experience.
- Led 12 software developers and 3 QA engineers on a project, and overall, in my career, **led 20+ software engineers and 5+ QA engineers**.
- Post-graduated from one of India's top ten technical universities in **Data Science and Engineering**.

Skills

- Certified in **AWS Machine Learning Specialty** and Azure AI fundamentals. Extensive experience working with AWS cloud. Also worked on IBM Cloud, and Digital Ocean.
- Well-versed with Python programming language, active volunteer, and one of the organizers of the PyCon India 2023.

ML tools	Grafana, LogStash, Databricks, MLFlow, LangChain, LangGraph, LangSmith, Kibana,
Web Frameworks	Flask, FastAPI, Django, ReactJS, Go Fiber, Gin
Programming Languages	Python, SQL, Golang, Javascript, Typescript, Shell script, Batch script, VB Script, COBOL, C, C++, MATLAB
Databases	MySQL, PostgreSQL, AWS ElastiCache, MongoDB, Redis, Elasticsearch, AWS Open search, IBM DB2
CI/CD Tools	Git, GitHub, Docker, Kubernetes, Jenkins, GitHub Actions, Cloudflare Pages, Terraform
Documentation Tools	Markdown, GitHub pages, Jekyll, Confluence, Jupyter Notebook, GitHub Wiki, Grammarly, Excalidraw, MS Visio, D2
Other Tools	Kafka, Spark, Nifi, Airflow, Nginx, Envoy proxy, Postman

EXPERIENCE

SEP 2021 – Present

Lead Systems Engineer/Lead MLOps Engineer
Freshworks

- **Designed end-to-end** LLMOps platform for the existing LLM solutions.
- Implemented solutions for **Model Monitoring, Logging, Evaluation, Versioning, Usage, and Adoption tracking**.
- Developed and **Designed a Dashboard** for monitoring the models' availability, infra usage, latency, efficacy, and throughput.
- **Standardized** Creation, Deployment across different production regions in various modes, Scaling up/down, and triggering retraining of the Model services.
- 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 (> 1.5TBs) MySQL queries, scaled up Database bottlenecks using sharding/partitions and reduced costs by removing old records.
- Other than the unit and sanity tests, **performance test**ed 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.

Nov 2023 – Present

AI Consultant
Independent

- 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.
- Developed **LangChain-based RAGs** to ingest client data and answer queries on top of it with streaming responses like ChatGPT.
- Coordinated with Sales team, Non-AI team to translate technical requirements and helped the CEO sign deals.
- Designed and implemented a voice bot to answer generic queries from customers and redirect leads to a CRM tool.
- Mentored team on best practices for maintaining a code, scaling, and designing multi-tenant services.

Apr 2023 – Present

Core Team Member
OdiaGenAI (Not for Profit)

- Developed an **instruction-tuned LLM**, Olive, for the Odia language.
- Collected and prepared **millions of Odia monolingual corpus**, which was used to prepare instruction tuning for a base LLM.

Dec 2018 – Aug 2021

AIOps Platform Lead
IBM

- Led two developers and one QA engineer and gained domain knowledge on AIOps and its use cases.
- 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.
- Designed and implemented multi-tenant architecture.
- Architected and implemented end-to-end **user authentication and authorization** in the ELK stack.
- Designed and proposed multiple transformation options for a **unified (real-time + batch) ingestion** and analytics framework for high-volume and real-time data analytics, providing end-to-end capabilities.

Oct 2016 – Dec 2018

Tech Lead
Wipro

- Led a team of 10 developers and QA engineers to develop an enterprise-grade search engine from scratch.
- Deployed the product **from PoC to a scalable production level** for 1000 concurrent users.
- Reduced the query time **from > 5 seconds to 100 ms**.
- Converted natural language queries (NLQ) to structured query languages (SQL) for structured data searches.
- In another European banking client project, entities were extracted from the email conversation between the client and the vendor using NLP algorithms. This automation helped the client reduce the headcount of 50 agents.

May 2015 – Sep 2016

Application Development Analyst
Accenture

- Automated complex human screen entries using VB Script and **saved worth $40,000+ per year** time of maintainers.
- Analyzed 10+ sev-1/2 production issues and built fixes.
- Worked with one of the top 3 banks in the world and gained immense domain knowledge on money transfer.
- Gathered and broke down the requirements into manageable problems to design solutions.
- Conducted the Unit Test, Sanity Test, and System Integration Test.

Jun 2012 – May 2015

System Engineer
Tata Consultancy Services

- I have worked with one of the top stock brokerage firms in the USA. I learned about mutual funds and stocks.
- I studied the BRD and HLD to understand the business, technical, and functional requirements.
- I organized and participated with the team to create the strategy for the enhancement and develop the optimum plan.
- I guided and delegated tasks to my junior associates.

EDUCATION

- Master of Technology in Data Science and Engineering, BITS, Pilani, India (2021-2023)
  - Thesis: LLM Alignment
- Bachelor of Technology in Electrical Engineering, SOA University, India (2008-2012)

LANGUAGE

- English – Fluent
- German – A1 certified

PERSONAL SKILLS

- Maintainer of an open-source PyPi package, OpenOdia, with 40,000+ downloads.
- Started a couple of volunteering teams, Odias in AI and OdiaGenAI.
- Led multiple volunteering teams: PyCon India 2023, Odia Wikipedia, and Mozilla Common Voice for Odia.
- Developed Shabdarasa, a wordle game for the Odia language with 100+ DAU.
- Actively volunteering on OpenStreetMap and other education-based causes.