Harshit **Singhai**

Software Engineer

(+91) 9560948115 | harshitsinghai77@gmail.com | fictionally-irrelevant.vercel.app | harshitsinghai77 | harshitsinghai

Work Experience

Deloitte Consulting USI

Gurugram, India

Backend Engineer Consultant - AI & Data — *Feb. 2022 - Ongoing*

- Collaborated with a leading US asset management firm to architect cloud analytics, APIs and backend data processing solutions on AWS
- Developed a **RAG** app with **AWS Bedrock** LLM, integrating data ingestion, embeddings, and vector store for fast, accurate content retrieval
- Designed **multi-agent** using **Bedrock & Strands SDK** workflow to auto-enrich AWS Glue Catalog metadata, cutting manual tagging by 60%
- Engineered serverless POC using AWS to convert Jira stories to ready-to-review GitHub draft PR, **cutting feature delivery time by 50%**
- Optimized FastAPI and PostgreSQL performance, reducing latency by **40%** and improving response times by **35%**, enhancing user satisfaction
- Eliminated inter-job dependency failures and mismanaged cron jobs by orchestrating cross-account data pipelines with AWS Step Functions
- Architected secure **PII infrastructure** with masking and segregated storage, enhancing compliance, saving **$50k/year in potential fines**
- Saved **10+ hours** weekly by migrating auditor reporting to AWS, automating data aggregation, and streamlining data extraction and reporting
- Built event-driven AWS Lambda pipelines triggered by S3 uploads to process, transform, and load financial files into Postgres and Athena
- Improved data availability with robust pipelines, achieving **99.8% uptime** and efficient ingestion from 7+ data streams with Glue and PySpark
- Migrated legacy data systems to AWS cloud, leveraging ECS, S3, and RDS, reducing infrastructure costs by **25%** and improving system scalability
- Reduced cost by **40%** using S3 Intelligent Tiering, replacing EMR with Glue autoscaling, S3 lifecycle policies, and Athena query optimization
- Implemented Glue data quality rules for data governance, automating anomaly detection to ensure reliable data for downstream systems

Quantive

Indore, India

Software Engineer - Platform — *Aug. 2021 - January 2022*

- Optimized microservices leading to **30% reduction** in memory usage & processing time through memory profiling & improved code readability
- Enhanced Pandas pipelines with vectorized operations and memory-efficient data types reducing processing time by **10%** for large datasets
- Increased data throughput by **40%** using Python Celery **producer-consumer architecture** for async processing and ClickHouse for storage

Deepsource

Bengaluru, India

Software Engineer - Machine Learning — *December 2020 - June 2021*

- Developed a recommendation system using XGBoost regression to rank code issues by likelihood of being resolved based on historical patterns
- Created 20k+ personalized ML models, one for each repository, to predict high-impact code issues, saving developers 10 hours/week
- Built adaptive ML retraining pipelines to prevent model decay, updating ML model only when new data improved precision and relevance

Programming Skills

Server Side	Python, JavaScript (React, Node.js), FastAPI, REST API
Data Engineering	PySpark, Pandas, AWS Glue, Airflow, S3, Athena, Hive, Presto, SQL
Machine Learning	Strands SDK, Qdrant, Bedrock KnowledgeBases, AgentCore, Langfuse, MCP, Multi-Agent, Bedrock, Scikit-Learn, Keras, MLOps
Database	SQL/NoSQL, PostgreSQL, MongoDB, AWS DynamoDB, Redis
Other	AWS, Docker, Git, Problem Solving, Data Structures Algorithms

Projects

- https://nemo-ai.netlify.app/: **NemoAI** - Cut Your Feature Delivery Time in Half. Nemo AI converts your Jira Stories into ready-to-review PR
- fictionally-irrelevant.vercel.app/posts/side-projects-2023: Shipped and deployed **20+** user-facing side projects, reaching over **1k users** globally
- https://fictionally-irrelevant.vercel.app : **Authored 72+** in-depth technical blog posts, delving into diverse areas like **AI, Cloud, Data, and web3**
- **MLH Fellowship**: Selected among the top **144** out of 30,000 applicants; contributed to **BentoML** open-source Python project

Accomplishments & Awards

2023	**Cohort Lead**, Deloitte AI Academy Explorer Program: Guided 80+ practitioners on Datacamp ML track	*Deloitte USI*
2022	**Applause Award**, recognized by Senior Leadership for delivering outstanding client outcomes	*Deloitte USI*
2019	**Mentored**, sophomore students for Python: Conducted and managed 30+ labs sessions and tutorials	*Bennett University*
2019	**National Finalist**, Smart India Hackathon: Enhanced Kotak Mahindra Bank's virtual assistant Keya	*Bhopal*

Education

Bennett University

Delhi NCR, India

Bachelor of Technology in Computer Science and Engineering (CSE) - 7.98 GPA — *Aug. 2016 - Aug. 2020*

- Main coursework: **Data Structures and Algorithms, Artificial Intelligence, Machine Learning, Cloud Computing, Big Data**