ANSH SHARMA

Data Engineer | ML Engineer | Data Scientist

+91 8368065604 | New Delhi, India

anshsharma2903@gmail.com | Portfolio | LinkedIn | GitHub.

PROFESSIONAL SUMMARY

MCA student specializing in Data Engineering and ML, with hands-on experience building LLM-powered platforms and NLP pipelines. Seeking opportunities in data engineering or applied ML roles.

PROJECTS

Spore — Local-First AI Data Platform | Jan 2026 – Present | WEB-PAGE | GITHUB

- Open-source, Dockerized Data-platform targeting **privacy-sensitive personals**, supports **local and cloud LLMs**.
- **Multi-source Data connectivity** (DB, Warehouse, API) with support for **SSH tunnels and SSL/TLS**.
- Memory-safe data-ingestion pipelines utilizing **columnar data formats** and chunking methodology to handle **15-20 GB+ datasets** in **Memory constraint environments**; eliminates **OOM** errors. Validated on a **6.7 GB / ~ 8M-row TSV dataset** ingested to PostgreSQL with end-to-end Data workflow on **pre-alpha** version.
- Strict **human-in-the-loop** execution – LLMs generate, humans review/edit and run; WebSocket endpoints **CORS-locked**, credentials and session data stored in a Dockerized **Redis instance** with Fernet encryption.
- Custom Jupyter-compatible **notebook** integration utilizing **sandboxed kernel instance** – mitigating **XSS and RCE risks**; supports native 3D interactive plots and extended tooling for EDA, ML and reporting.

Wingman — Open-Source GitHub Copilot Alternative | 2024 | GITHUB

- Built a VS Code extension leveraging **LLMs** to provide context-aware code completions, documentation generation, test-case creation, and real-time debugging suggestions.
- Integrates local model serving with **Ollama** for faster inference and fully privacy-focused code-assistance – **no data leaves** the user's machine.
- Tuned prompt engineering flow for real-time documentation and improvement suggestions.

Digital Asset Management System | 2023

- Designed and developed a production-ready web application for centralized digital asset management, enabling organized storage and retrieval of organization assets.
- Implemented secure server-side **Google OAuth 2.0** authentication for user identity and access control.
- Deployed client on **Vercel** and backend on **Render**, ensuring high availability, scalability and CI/CD-friendly deployment workflow.

ACADEMIC PROJECT

Data Annotation and Collection Project | 2025

- Collected **20k+ code-mixed Hinglish data records** from **Reddit, Twitter, Facebook, Instagram, YouTube**, etc. using **Selenium/Selenium-stealth, APIs, public JSON endpoints**. Preprocessed raw data for NLP workflows including **data annotation** using Excel and **Label Studio**, contributed to pipeline optimizations.
- **Mentored** 4 MCA/MTECH students **across institutions** on data collection techniques and methodology.
- Independently proposed an **automated annotation pipeline**: weak model pre-annotates samples with confidence scores and routes them – auto-accept (>0.9), human review (0.6-0.9), full annotation (<0.6) – clustered similar records to reduce cognitive load during manual annotation – Additional lightweight TTS model proposal to reduce reading fatigue.

SKILLS

Languages: Python, R, SQL, NoSQL

ML/Data: Scikit-learn, Pandas, Numpy, Apache Arrow, Apache Superset, RAG, Vector Databases

LLM/NLP: Ollama, LM Studio, NLTK, LLM Integration,

Infra: Docker, Flask, Git, REST APIs, WebSockets

Databases: PostgreSQL, MongoDB, Supabase, Redis, DuckDB

EDUCATION

Master of Computer Applications — **8.1/10 CGPA**

Guru Gobind Singh Indraprastha University, Jagan Institute of Management Studies — *08/2024 – Ongoing*

New Delhi

Bachelor of Computer Application — **7.9/10 CGPA**

Guru Gobind Singh Indraprastha University, Maharaja Surajmal Institute — *11/2021 - 07/2024*

New Delhi