Shubham Pareek

MS in Data Science Candidate | Data Science, Machine Learning, and AI Internships

sp4553@columbia.edu | linkedin.com/in/shubhampareek1211/ | github.com/shubhampareek1211 | Portfolio

Education

Columbia University New York, NY
Master of Science in Data Science *Expected Dec 2026*
Teaching Assistant – Cloud Computing, Introduction to Databases

SRM Institute of Science and Technology Chennai, India
Bachelor of Technology in Computer Science and Engineering *May 2023*

Skills

Languages: Python, SQL, R, C++
ML/AI: PyTorch, scikit-learn, Transformers, RAG, LangChain, Prompt Engineering, LoRA/PEFT, Time-Series Forecasting, Statistical Modeling, Deep Learning
Data / Analytics: PostgreSQL, dbt, Tableau, Power BI, Excel, Spark, Matplotlib
Cloud / DevOps: Google Cloud Platform (GKE, Compute Engine), AWS RDS, Git, GitHub, FastAPI, n8n
Certifications: Google Cloud Associate Cloud Engineer; Google Cloud Professional Cloud DevOps Engineer

Experience

Student Research Worker, CUIMC – New York, NY Mar 2026 – Present
- Built a Python pipeline to generate and de-identify synthetic EHR data by combining patient-level temporal shifting with HIPAA-compliant PHI redaction across 18+ entity types, creating privacy-preserving clinical text for downstream ML and GenAI workflows.
- Deployed de-identified clinical datasets to AWS RDS and implemented SSO-based, role-based access controls, enabling secure and governed access for research teams.

Data Analyst Intern, Weatherhead East Asian Institute, Columbia University – New York, NY Sep 2025 – Present
- Built automated Python ETL pipelines to clean, transform, and validate financial market datasets with 99%+ processing accuracy, improving reliability of downstream analysis and reporting.
- Developed PostgreSQL data models and dbt-powered transformation workflows to support SQL analytics, trend analysis, benchmarking, and forecasting across 6+ months of historical data.

Programmer Analyst, Cognizant Technology Solutions – Hyderabad, India Jan 2024 – Jul 2025
- Resolved production issues for Google Cloud enterprise customers across GKE clusters, compute infrastructure, and data pipelines, reducing MTTR by 40% and maintaining 98% customer satisfaction.
- Performed root-cause analysis on cloud infrastructure and pipeline failures, improving system reliability and strengthening incident response across cross-functional support teams.
- Contributed to the development and deployment of an internal CRM platform on GCP, helping streamline internal service workflows and improve operational efficiency.

Projects

Resume Griller: AI-Powered Interview Simulator GitHub
Python, TypeScript, LoRA, ChromaDB, RAG
- Fine-tuned an open-source LLM with LoRA/PEFT on interview dialogue data to build an AI interview simulator capable of follow-up question generation, response scoring, and gap identification.
- Built a RAG pipeline using ChromaDB and sentence-transformer embeddings with semantic resume chunking and low-latency retrieval, enabling grounded, resume-aware interview interactions.

Electricity Forecasting GenAI Agent GitHub
Python, Time Series, XGBoost, LightGBM, FastAPI, n8n
- Built a large-scale forecasting pipeline on 13M+ electricity observations across 368 clients, using clustering, lag-based features, and machine learning models to predict multi-seasonal demand patterns.
- Developed an LLM-powered agent in n8n that translated natural-language forecasting requests into API calls and generated automated forecast summaries for business-facing consumption.