

ABDUL RAFAY

Senior Full-Stack Engineer | Python, Node.js, Vue | Generative AI

Islamabad, Pakistan | rafay3515@gmail.com | +92 334 5162516 | linkedin.com/in/rafay-ah | rafay-ah.github.io | github.com/rafay-ah

SUMMARY

Full-stack engineer with five years in production, building products from the first requirement to paying users. At DataGroomr, a Salesforce data-quality platform used by around 1,500 companies, leading a team of seven and building its AI products: an in-app assistant that diagnoses data problems and acts on them, and an enrichment engine that completes CRM records from 25+ data providers. Before that, high-volume Python backends and applied machine learning: document and payment pipelines for fintech clients, Arabic-dialect NLP over 100M+ social posts.

EXPERIENCE

Senior Full-Stack Engineer, DataGroomr

Mar 2025 - Present

Remote, US | Salesforce data-quality platform (duplicate detection, cleansing, verification, enrichment); multi-tenant SaaS on AWS; around 1,500 B2B customers

- Lead a team of seven engineers day to day: run stand-up, break epics into work and distribute it, review every pull request in the group, manage release branches and hotfixes, and act as the single point of escalation between customer support and engineering.
- Built the DataGroomr Assistant, an in-app AI agent for CRM data, live for every customer. It diagnoses problems in a customer's own Salesforce data, explains why records matched, drafts rules from plain English, and takes actions such as merging duplicates behind explicit confirmation. Python, FastAPI and LangGraph: 16 intent-routed flows over 38 tools, RAG over pgvector, SSE streaming into Vue, voice mode.
- Built the Enrichment engine that lets sales and marketing teams complete CRM records with one plain-English prompt. An AI planner turns the prompt into an execution plan; 25+ data providers (ZoomInfo, D&B, Apollo, live web search) are integrated through MCP and OAuth; a streaming Node.js pipeline enriches millions of records and writes results back to Salesforce, in bulk or in real time.
- Added AI to the platform's rule builders: non-technical admins describe a data-quality, transformation or merge rule in plain language and get a working rule, an explanation of an existing one, or an AI review of thousands of duplicate groups in one run. Support requests about rule authoring fell by roughly 90%.
- Delivered real-time duplicate detection on Salesforce record changes with automatic merging, including the dataset-level locking that keeps live events, auto-merge and scheduled analyses from colliding, so duplicates are caught the moment they appear instead of in periodic clean-ups.
- In the Python matching engine, built similarity matching (MinHash + LSH) that finds fuzzy duplicates across datasets of 10M+ records, an information-theoretic method for choosing which field combinations are worth comparing, and match-confidence breakdowns that show users why two records were paired.
- Built the per-tenant Postgres reporting layer behind the analytics dashboards (BullMQ materializers and backfill workers, Vue 3 front end, churn-risk signals) and the Stripe subscription flow with monthly credit provisioning.
- Own the platform's LLM infrastructure: a model-routing gateway across Anthropic, OpenAI, Gemini and AWS Bedrock (including customers' own keys), prompt caching and retry policies, tracing and observability in Langfuse, and an LLM evaluation suite (golden cases, LLM-as-judge) that catches prompt and model regressions before release.

Python / ML Engineer, Sila Insights

Mar 2024 - Mar 2025

Dubai, UAE (Remote, contract) | AI consumer-intelligence platform for brands in the Middle East; data and research team

- Designed and built the social-data collection service behind Sila's brand and influencer analytics: an async FastAPI orchestrator over Apify collectors for Facebook, Instagram, TikTok, X, Google Reviews and TripAdvisor, with de-duplication, per-platform normalization and Terraform-managed Azure infrastructure. Ingested over 100 million posts and reviews.
- Trained and served the Arabic NLP models (dialect, sentiment, emotion, irony) across 20+ dialects behind BrandHealth360, the flagship brand-tracking product, on Redis-queued FastAPI workers; training and deployment automated with Docker and Airflow.
- Built the LLM layer over that data: a multi-tenant RAG pipeline (Pinecone, Chroma, LangChain agents) with structured-output guardrails feeding dashboards and CRM integrations. Cut monthly inference spend by about 40% through batching, caching and model-tier routing.

Python Engineer, Neural Lines

May 2022 - Feb 2025

Islamabad, Pakistan | Backend, ML and document-intelligence products for fintech, payroll and media clients

- Built a document-processing platform for Soluma: multi-language OCR, barcode reading, PDF splitting and field extraction, delivered as containerized FastAPI services that customers integrate directly.
- Built the ingestion pipeline that loaded millions of historical payment orders for a Slovenian bank (FastAPI, Celery, Redis): parsing, normalization, cross-checks against internal databases, structured output for reconciliation.
- Built the document-intelligence backend of Spend Console, a payroll and spend-automation product: invoice extraction on Azure Form Recognizer with validation and classification at production scale, plus spend-categorization models retrained from user corrections.
- Built the live-vision system for a Red Bull-sponsored Fortnite tournament: OCR over seven parallel screen feeds detected eliminations in real time and switched the broadcast to the eliminating player, at about 99% accuracy in production.
- Early LLM product work in 2023: an exam-simulation system on the GPT API and LangChain that generates questions, grades free-form answers against rubrics and adapts difficulty per student.

Associate Android Developer, O3 Interfaces

Feb 2022 - May 2022

Islamabad, Pakistan

- Kotlin Android development from design through Play Store beta release.

SKILLS

Languages: Python, JavaScript, TypeScript, SQL

Backend: Node.js, Express, FastAPI, Flask, REST APIs, WebSockets, Server-Sent Events, microservices, event-driven architecture, distributed systems, streaming pipelines with backpressure, BullMQ, Celery, Redis, multi-tenant SaaS, OAuth 2.0

Frontend: Vue 3, Vuex, Pinia, Vite, webpack, Tailwind CSS, Socket.IO

Data: MongoDB, PostgreSQL, Redis, pgvector, vector databases (Pinecone, Chroma), Supabase

Generative AI: Anthropic Claude, OpenAI, Gemini, AWS Bedrock, LLM agents and tool calling, Model Context Protocol (MCP), LangGraph, LangChain, RAG, embeddings, prompt engineering, prompt caching, LLM evals (promptfoo, LLM-as-judge), observability (Langfuse)

Machine learning: PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, NLP, OCR (Azure Form Recognizer, Tesseract), computer vision, record linkage (MinHash, LSH)

Cloud and delivery: AWS (ECS Fargate, Lambda, S3, Bedrock, CloudWatch), Azure, Cloudflare Workers and Pages, Docker, Terraform, CI/CD, Salesforce APIs, Stripe, Paddle

Practices: System design, architecture and code review, unit and integration testing (Mocha, pytest, Vitest, Playwright), agile delivery, mentoring, AI-assisted development (Claude Code)

INDEPENDENT PRODUCTS

- **FilerKaro (filerkaro.com), 2026.** Self-serve tax-filing product for Pakistani freelancers that turns payment-platform exports into a ready-to-file return. Designed, built, launched and marketed solo: Vue 3, TypeScript, Express, Supabase, Paddle, Cloudflare.
- **ExpiryOps, AnswerRank and AgentReady, 2026.** Three SaaS tools (domain and certificate expiry monitoring for agencies; tracking how AI assistants recommend a business; AI-agent readiness checks for websites), each taken from idea to live product.

EDUCATION AND CERTIFICATIONS

B.Sc. Computer Science, COMSATS University Islamabad: 2018 - 2022, AI and machine learning specialization

Data Science Specialization: Johns Hopkins University

Deep Learning Specialization and TensorFlow Developer Certificate: DeepLearning.AI

COMMUNITY LEADERSHIP

- Lead trainer for a TechLift MERN-stack bootcamp, the national industry-led program launched by the Ministry of IT and the Pakistan Software Export Board; trained a cohort of about 40 graduates.
- Ran data analytics workshops (Python, NumPy, pandas, statistics and analysis methodology) at Taylor's University, Malaysia, in 2024 and 2025.
- Delivered a five-day onsite training on automation testing and CI/CD pipelines at AKSA SDS.