Last updated in Jun 2025

Subhayu Kumar Bala

Kolkata, India ✉ balasubhayu99@gmail.com ☎ +91 93828 77751 subhayu-kumar-bala subhayu99

Intro

Data and infrastructure engineer with 3+ years of experience bridging traditional data engineering with modern AI systems, specializing in Python, SQL, and cloud platforms (AWS/Azure/GCP).

Track record of exceptional performance optimization, including reducing a 27-hour SQL procedure to 5 seconds by implementing innovative Python-based solutions for processing 80M+ rows of data.

Expert in building scalable architectures through event-driven design, containerization, and agentic LLM systems, consistently delivering solutions that transform complex business requirements into technical reality.

Technologies

Core Programming: Python, SQL, BASH, JavaScript

Data Engineering & Processing: Pandas, NumPy, DuckDB, PySpark, ETL/ELT Pipelines, DBT, Data Modeling

Data Orchestration & Workflow: Airflow, GCP Workflows, Kafka, Treasure Data, Microsoft Fabric, ADF

AI & LLM Engineering: OpenAI APIs, Gemini, Agentic Architecture, MCP, A2A, LangChain, Axolotl, LLM Integration, Prompt Engineering, RAG, SerpAPI, LLM Fine-tuning, TensorFlow, Keras, SKLearn

Cloud Infrastructure: AWS, Azure, GCP, S3, Terraform, Pulumi, ARM Templates

DevOps & CI/CD: Docker, Kubernetes, Azure DevOps, CI/CD, Linux, NGINX, Jenkins, Git, Message Queues

Databases & Vector Stores: MongoDB, PostgreSQL, MySQL, MS SQL Server, BigQuery, Chroma, FAISS, Neo4j

API & Application Development: FastAPI, Django, DRF, Flask, Typer CLI, REST APIs

Data Visualization: Looker Studio, PowerBI, Matplotlib, Streamlit, Plotly

Specialized Technologies: Web Scraping, CDP, SFMC, Quantum Computing Simulation

Experience

FiftyFive Technologies, Data Engineer — Gurugram, India (Remote)
Jun 2022 – present

- **Data Engineering & Analytics:** Built ETL/ELT pipelines processing millions of healthcare and marketing records using Python, SQL, PySpark, and cloud platforms (AWS/Azure/GCP), with expertise in performance optimization achieving 99.8% runtime improvements.
- **AI/LLM Integration:** Developed production AI systems using OpenAI APIs, LangChain, and fine-tuned models (Mistral, Llama) for document processing, outreach automation, and agent-based architectures with 95%+ accuracy.
- **Backend, Cloud Infra & DevOps:** Managed containerized deployments with Docker/Kubernetes, implemented CI/CD workflows, and built FastAPI applications serving 100k+ users with real-time dashboards and REST APIs.

FiftyFive Technologies, Software Engineer Intern — Gurugram, India (Remote)
Jan 2022 – May 2022

- **Backend Development:** Built Django/DRF applications with PostgreSQL/MongoDB databases, developed REST APIs, and created automation tools for cable network management using geospatial data processing.
- **DevOps Implementation:** Established Azure DevOps CI/CD pipelines for platforms serving 50k+ daily users, managed database migrations, and integrated third-party APIs for operational automation.

Education

CIEM, (B.Tech - Information Technology) — Kolkata
Jul 2018 – Jun 2022

- CGPA: **8.57**/10 (Certificate)

Selected Projects

Johnson & Johnson - Healthcare Data Platform Feb 2025 – Jun 2025

- Architected and maintained scalable data pipelines (Bronze-Silver-Gold layers) in Treasure Data using SQL, processing millions of records for JP and ANZ Healthcare Professionals' marketing analytics.
- Engineered complex data transformation logic including a 12-scenario truth table for multichannel consent processing, integrating data across platforms (CDP, S3, Treasure Data, SFMC, GA4) using Python and APIs.
- Ensured enterprise-grade data quality and pipeline reliability by implementing robust consistency checks and automated data processing workflows for regulated healthcare environments.

Wade Insight - Cloud-Native Data Orchestration Platform Nov 2024 – Feb 2025

- Enhanced core pipeline orchestration features for WADE's Azure SaaS platform, implementing comprehensive continue_on_failure mechanisms with SQL dependency handling and advanced job execution tracking.
- Led migration of enterprise data pipelines from Azure Data Factory to Microsoft Fabric Data Factory, managing ARM template adaptation and deployment for clients processing millions of records.
- Developed automated healthcheck processes leveraging Azure APIs to monitor running pipelines, reducing manual troubleshooting overhead by 80% and enabling faster insight delivery.

Prospexs - AI-Powered Outreach Platform Jul 2024 – Oct 2024

- Built an AI-powered outreach platform using Python/FastAPI and MongoDB, reducing manual prospecting effort by 60% through intelligent automation.
- Integrated multiple APIs (OpenAI, Perplexity, LinkedIn) to validate profiles and generate personalized communications, improving client response rates by 45%.
- Developed email generation system with dynamic tone adjustment capabilities, enabling scalable personalized outreach for B2B sales teams.

QxLab - State-of-the-Art Agent-Based LLM System Jan 2024 – Jun 2024

- Built a production-ready agent-based LLM system using fine-tuned Mistral7B and Llama 13B models for real-time API data processing, achieving over 95% accuracy with custom schema management for tool orchestration.
- Developed advanced CLI tool with Typer for language model fine-tuning and dataset manipulation, processing millions of data points and 10B+ tokens in minutes through optimized pipelines.
- Orchestrated FastAPI deployment on Docker for GPU-accelerated model inference with multi-threading capabilities, enabling scalable AI model serving infrastructure.

CV Advisors - High-Performance Financial Data Processing Jan 2024 – Jan 2024

- Developed a Python/Pandas/DuckDB proof-of-concept for financial data processing pipeline, reducing runtime of a 1900+ line SQL procedure from 27.5 hours to under 5 seconds.
- Demonstrated Python's capability for high-performance data processing on 80M+ rows across 150 clients, enabling batch execution and significant operational efficiency improvements.

LoopKitchen (now Loop) - Food Delivery Intelligence Platform Sep 2022 – Jun 2023

- Built core data infrastructure from scratch for a food delivery analytics startup using comprehensive GCP stack: FastAPI/SQLModel APIs on Cloud Run/Functions, BigQuery for data warehousing, Firestore for metadata, orchestrated via GCP Composer and Workflows.
- Developed multi-platform data ingestion pipelines scraping and processing millions of orders from UberEats, DoorDash, and Grubhub APIs, enabling real-time performance analytics for restaurant brands and franchises.
- Created custom Streamlit monitoring dashboard providing real-time visibility into complex orchestration workflows with granular tracking (brand→region→chain→store→order level), essential for debugging long-running processes across multiple third-party platforms.
- Contributed as core team member (10-person startup) working directly with technical leadership, helping build the foundation for a platform that secured $6M Series A funding and now serves major restaurant chains like Dave's Hot Chicken and Freddy's.

Other Notable Projects

Logical Contract (Oct 2023 - Dec 2023): Implemented an AI-powered system for startups, generating tailored employment agreements and a legal chatbot for inquiries.

SlideNinja (Jul 2023 - Oct 2023): Developed and deployed an AI-powered RAG platform for a McKinsey partner, leveraging GPT-3.5, LangChain, and SerpAPI to generate presentations from minimal input during the early GenAI boom.

Eningo (Apr 2022 - Aug 2022): Developed a cloud-based platform automating cable network issue requests with geospatial data processing and Postgres to MongoDB migration for enhanced scalability.

NIBE (Feb 2022 - Mar 2022): Managed CI/CD pipelines on Azure DevOps, impacting over 50,000 daily users.

Personal Projects

DatasetPipeline May 2025

- Developed a production-ready CLI tool for transforming messy datasets into ML-ready formats; supports SFT, DPO, semantic deduplication, and quality analysis with plugin architecture.
- Features smart role mapping, auto-formatting for OpenAI-style training, and reproducible workflows via YAML/JSON configuration for enterprise ML pipelines.
- Published on PyPI ↗ and open-sourced on GitHub ↗ with extensible architecture for custom loaders, formatters, and analyzers.

Smart Commit May 2025

- Built an AI-powered CLI tool using Python/Typer that generates context-aware git commits via OpenAI or Anthropic models, enhancing developer productivity and project history quality.
- Engineered deep repository analysis of tech stacks, commit patterns, and file changes with flexible global/local .toml configurations for project-specific conventions.
- Published on PyPI ↗ and GitHub ↗ with Model Context Protocol (MCP) server for direct AI assistant integration.

Publication

QuDiet: A Classical Simulation Platform for Qubit-Qudit Hybrid Quantum Systems Mar 2023

Subhayu Kumar Bala, Turbasu Chatterjee, Arnav Das

10.1049/qtc2.12058 ↗ (IET Quantum Communication)