Subhayu Kumar Bala

balasubhayu99@gmail.com +91 93828 77751 subhayu99.github.io subhayu-kumar-bala
subhayu99

Intro

A Data & Infrastructure Engineer with 3+ years of experience building high-performance, intelligent systems that drive measurable business outcomes. I specialize in turning complex data problems into efficient, scalable software.

I leverage deep expertise in Python, SQL, and multi-cloud architecture (AWS/Azure/GCP) to deliver transformative results, from slashing a 27-hour data process to just 5 seconds to engineering production-grade agentic LLMs.

My work has directly contributed to critical business milestones for clients, including enabling a startup to secure a $6M Series A funding round and ensuring enterprise-grade data compliance for regulated industries.

Technologies

Core Languages & Tools: Python, SQL, BASH, JavaScript, Git

Data Engineering & Orchestration: ETL/ELT Pipelines, Data Modeling, PySpark, Pandas, NumPy, DuckDB, DBT, Airflow, GCP Workflows, Kafka, RabbitMQ, Microsoft Fabric, ADF, Treasure Data, GeoPandas, PostGIS

AI & LLM Engineering: OpenAI APIs, Gemini, LangChain, Agentic Architecture (MCP, A2A), RAG, Axolotl, LLM Fine-tuning (SFT, DPO), Prompt Engineering, Semantic Deduplication, TensorFlow, Keras, SerpAPI

Cloud & DevOps: AWS, Azure, GCP, S3, Docker, Kubernetes, Terraform, Pulumi, ARM Templates, CI/CD Pipelines, Jenkins, Azure DevOps, NGINX, Linux, Cloud Cost Optimization

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

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

Security & Quality Assurance: Zero-Trust Architecture, Cryptography (E2EE, AES-GCM, RSA), IAM, OAuth2/JWT, SSO/SAML, RBAC, Secret Management, Pytest, Unittest, Pydantic, Pandera, Great Expectations

Specialized & Visualization: CDPs, SFMC, Web Scraping, Looker Studio, PowerBI, Streamlit, Plotly, Matplotlib, Quantum Computing Simulation, Agile (Scrum/Kanban), Code Reviews, Technical Consultation & Solution Design

Experience

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

- **Delivered >99.99% Performance Gains:** Re-architected a 27.5-hour legacy SQL procedure into a 5-second Python/DuckDB process, successfully handling 80M+ financial records and unlocking significant operational efficiency for the client.
- **Engineered and Shipped Production AI:** Built and deployed a production-grade agentic LLM system from the ground up, using fine-tuned Mistral/Llama models to achieve >95% accuracy in real-time data processing and tool orchestration.
- **Enabled a $6M Series A Funding Round:** Architected the foundational data platform from scratch on GCP for a high-growth startup, providing the critical infrastructure and real-time analytics that were instrumental in their successful funding.
- **Ensured Enterprise Data Compliance:** Implemented robust data pipelines for a major healthcare client (Johnson & Johnson), engineering complex consent logic to meet strict data governance and privacy standards in regulated markets.

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

- Contributed across the full development lifecycle, from building a geospatial backend platform that earned direct client commendation to establishing CI/CD pipelines on Azure DevOps for services supporting 50k+ daily users.

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 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.
- Intelligently adapts to different projects by analyzing the existing tech stack, commit history, and file changes, ensuring contextually relevant and consistent messages.
- Published on PyPI and GitHub with Model Context Protocol (MCP) server for direct AI assistant integration.

DocumentAccessPOC Jan 2025

- Designed and built a zero-trust secure document system to solve granular access control challenges where traditional RBAC/ACLs fail, ensuring data confidentiality even from system administrators.
- Implemented a robust cryptographic model featuring end-to-end encryption (AES-GCM) and secure key exchange (RSA) to enforce permissions at a data level, not just application logic.
- Built a FastAPI interface for secure document sharing and revocation, with the full project documented and open-sourced on GitHub.

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)