

Radha Parikh

[Website](#) | [LinkedIn](#) | [GitHub](#) | radha.parikh11@gmail.com | +1 (412) 596-0002

SKILLS

Data Engineering:	PySpark, Apache Kafka, Azure Databricks, ETL/ELT, Stream Processing, Event-Driven Architecture, Data Modeling
Cloud & Databases:	Azure (Data Lake, Blob Storage, Function Apps, DevOps), CosmosDB, PostgreSQL, MySQL, Redis
ML/AI:	PyTorch, TensorFlow, LLMs, Transformers, FAISS, Neural Networks, scikit-learn, XGBoost, MLflow, Model Deployment
DevOps & Automation:	Docker, Kubernetes, CI/CD (Jenkins, GitHub Actions), Git, Power BI, Tableau, Streamlit
Programming:	Python, SQL, Java, JavaScript

PROFESSIONAL EXPERIENCE

GoodFutures

Founding Engineer

Pittsburgh, PA

Sept. 2025 - Present

- **Designed** and implemented **PostgreSQL** database architecture with **normalized schema**, foreign key constraints, and strategic **indexing** on join columns, enabling iterative product development while maintaining data integrity and **query performance** for future scale
- **Built RESTful API** layer with comprehensive **input validation**, **error handling**, and structured **logging**, establishing **maintainable** code patterns that support rapid feature iteration as platform grows to serve increasing user load
- **Established CI/CD** infrastructure with **GitHub Actions** and **automated testing**, enabling rapid feature **deployment** while ensuring code quality and system reliability as team and codebase expand
- **Architected** flexible **data models** supporting matching algorithm requirements, implementing efficient **query patterns** and database **indexes** that allow for **algorithm evolution** without major schema refactoring

Fractal Analytics

Data Engineer

Bengaluru, India

Aug. 2021 – Jun. 2024

Transportation Analytics Platform for Fortune 50 CPG Company with 500+ users

- **Redesigned data ingestion** from full load refresh (4 hours) to incremental **CDC (15 minutes)** using **PySpark** and **CosmosDB**, implementing **upsert** logic based on order lifecycle events, carrier tendering, and LRDT status fields cutting cloud costs by 62.5% (**\$32K to \$12K monthly**)
- Architected **real-time** shipment monitoring platform processing **50,000+** weekly loads across **15+ retailers** with sub-5-second latency and **99.9% availability**, serving 500+ business users to **minimize OTIF** delivery failures
- Built **real-time data pipeline** feeding **risk-scoring** engine integrating **3 enterprise data sources** (SAP ERP purchase/sales orders, TMS carrier data, FourKites real-time GPS) via REST APIs, calculating lane-specific risk scores for **15,000+ shipments** daily with sub-5-second latency
- **Migrated** monolithic Azure Function Apps to distributed **Databricks** architecture with **PySpark**, implementing **parallel processing** for **100+ GB** daily data loads and reducing **batch processing time by 50%**
- Established **Azure DevOps CI/CD** pipelines with **automated testing** and code review workflows, reducing production incidents by **75%** while mentoring 3 junior engineers on distributed systems patterns

Healthcare Finance Forecasting Platform for a Dutch Multinational Company

- Built automated monthly **ELT** pipeline on **Databricks** processing 52 entity-currency combinations from Azure Data Lake to Blob Storage with automated error handling and recovery, maintaining **zero data loss** over 12-month production run
- Designed data validation framework with schema validation and anomaly detection for financial time-series data, implementing **automated checks** that caught **40%** of data quality issues **pre-ingestion**, reducing downstream analytics errors

Marketing Mix Modeling Platform for Fortune 100 Beverage Company

- **Engineered data validation** and quality framework for multi-dimensional marketing datasets (**10+ channels, 50+ brands, 3+ years historical data**) with schema validation and statistical anomaly detection, ensuring data integrity for downstream **TensorFlow models**
- Developed comprehensive unit test suite for **TensorFlow**-based simulation models using pytest, achieving **95%+ code coverage** with automated test data generation for brand covariance and hierarchical regularization components

EDUCATION

CARNEGIE MELLON UNIVERSITY, HEINZ COLLEGE

Master of Information Systems Management

Pittsburgh, PA

Aug. 2024 – Aug. 2025

Intro to Deep Learning (PhD), AI Venture Studio, Operationalizing AI, Distributed Systems, OOP in Java, Agile Methods, Digital Transformation

MANIPAL INSTITUTE OF TECHNOLOGY

Bachelor of Technology in Aeronautical Engineering, Minor in Data Science

Manipal, India

Aug. 2017 – Aug. 2021

PROJECTS

Real-Time Air Quality Prediction System [\[github\]](#)

Feb. 2025 – Apr. 2025

- Architected event-driven streaming pipeline using Apache Kafka processing 1,000+ events/second with producer-consumer pattern for real-time sensor data ingestion
- Built ML training pipeline with scikit-learn and XGBoost implementing feature engineering and cross-validation, containerized with Docker Compose and integrated MLflow for experiment tracking

Intelligent Scheduling & Estimation for Suffolk Construction

Jun. 2025 – Aug. 2025

- Developed NLP data processing pipeline automating schedule generation by mapping 1,000+ cost estimate line items to activities using TF-IDF and sentence transformers
- Built similarity scoring system with weighted ensemble approach, integrated with Label Studio for human-in-the-loop validation and continuous retraining

GitaGPT - Context-Aware Bhagavad Gita Recommendation System [\[github\]](#)

Jan. 2025 – Apr. 2025

- Built Android application with semantic matching NLP system recommending Bhagavad Gita verses, processing 500+ user queries during beta testing
- Designed client-server architecture with Apache TomEE backend for processing user prompts, deployed using Docker in GitHub Codespaces