

PRIYA GANESAN

+1-857-339-8452 | murugesanganesan.s@northeastern.edu | [linkedin.com/in/shanmuga-priya-m-g-6678001ba](https://www.linkedin.com/in/shanmuga-priya-m-g-6678001ba) | github.com/Priya1401 | Boston, MA - 02119

Data-centric AI Engineer specializing in scalable infrastructure for machine learning. Leverages 2 years of experience optimizing real-time ETL pipelines at Goldman Sachs to design scalable AI architectures and automate end-to-end learning lifecycles in production.

EDUCATION

- Northeastern University, Khoury College Of Computer Sciences** Sept 2024 - Present
Master of Science in Artificial Intelligence - CGPA: 3.6/4.0 Boston, United States
- Anna University, Sri Sivasubramaniya Nadar College Of Engineering** Aug 2018 - May 2022
Bachelor of Engineering in Computer Science - CGPA: 8.42/10 Chennai, India
 - Thesis - Developed a hybrid deep learning pipeline (U-Net + NLP) to assess post-disaster regional impact using satellite imagery & Twitter sentiment, performing **correlation analysis on heterogeneous datasets**.

TECHNICAL KNOWLEDGE

- Languages: Python, Java, Scala, SQL
- Data Engineering: Kafka, PySpark, Hadoop, Hive, HDFS, Elasticsearch
- MLOps & DevOps: Airflow, MLflow, Docker, DVC, GitHub Actions (CI/CD)
- Cloud & Infra: GCP, Google Kubernetes Engine (GKE), Terraform
- ML/AI: Scikit-learn, PyTorch, TensorFlow, Keras, BERT, LSTM, U-Net
- Backend/Web: FastAPI, Flask, REST APIs, ReactJS

EXPERIENCE

- Software Engineer - Goldman Sachs (Data Engineering Team), Bengaluru, India** July 2022 - July 2024
 - Engineered high-performance real-time ETL pipelines using Kafka and Elasticsearch, processing over **12 Terabytes** of daily financial data with **<30 seconds** latency for optimized HDFS loading.
 - Optimized Parquet storage architecture via manual compaction strategies, achieving a **23% improvement** in storage efficiency and significantly reducing downstream query latency for analytics teams.
 - Developed a PySpark-based automated validation framework with 32 custom constraints and PyTest integration, catching an average of **50+ critical data anomalies weekly** prior to production ingestion.
 - Managed the integrity of a complex distributed ecosystem spanning 230 data groups and 18K+ datasets, reducing data availability incidents by **30%** through proactive monitoring.
 - Implemented CI/CD pipelines via GitHub Actions to automate build, test, and deployment stages, accelerating release cycles from **weekly to daily** and eliminating manual deployment errors.
 - Onboarded 12+ new external datasets through custom web scraping and file-based ingestion workflows, transforming raw inputs into structured HDFS assets for dynamic reporting.
- Software Intern - Goldman Sachs, Bengaluru, India** June 2021 - Aug 2021
 - Developed an interactive ReactJS + Redux dashboard to streamline the creation and management of securities within the firm's internal database, improving data entry efficiency for analysts by **~40%**.
 - Designed and implemented secure RESTful APIs for real-time communication between UI and backend systems, supporting thousands of daily internal user requests.
 - Optimized front-end components for responsiveness and performance, reducing page load times by **over 30%** to ensure a smoother workflow for financial analysts.
 - Collaborated with senior engineers to integrate the application into the production environment, gaining hands-on experience with agile practices in large-scale financial systems.

PROJECTS

- AudioSEEK: Timestamp-Grounded Spoiler Free Q&A for Audiobooks** | Python, GCP (GKE), Terraform, RAG Sept 2025 - Dec 2025
 - Architected a multimodal RAG system transforming audiobooks into interactive Q&A; engineered a timestamp-alignment algorithm using **FasterWhisper** to ground LLM responses in specific audio segments, achieving **<10% WER** on validation sets.
 - Implemented a "Spoiler-Safe" reasoning engine using Google **Gemini** and **BGE-m3** embeddings, filtering future plot points to preserve narrative integrity with **>60% ROUGE-L** semantic consistency.
 - Deployed a scalable **MLOps** infrastructure on **GCP** (GKE) via **Terraform**, featuring horizontal auto-scaling and incorporating **MLflow** for experiment tracking and **FastAPI** for low-latency inference.
- LLMPedia: AI-Powered Knowledge Platform** | Python, LangChain, OpenAI API, Flask, ReactJS June 2025 - Aug 2025
 - Developed a context-aware knowledge platform leveraging large language models for information retrieval and conversational learning.
 - Implemented RAG-based querying with structured prompt templates and a full-stack web interface for visualization and dynamic response generation.
- Team-Play Archetypes in League of Legends Ranked Matches** | Python, K-Means, PCA, Matplotlib, Scikit-learn May 2025 - July 2025
 - Built K-Means clustering models on a dataset of ~68K ranked players to classify gameplay behavior using metrics such as fighting efficiency, farming, vision control, and objective participation.
 - Engineered player features, performed data cleaning and z-score standardization, validated clusters via silhouette and elbow analysis, and identified three archetypes visualized through PCA and radar plots.
- Twitter Sentiment Analysis** | ML, LSTM, BERT, LDA, Flask Feb 2025 - April 2025
 - Created a real-time sentiment classification system comparing ML models (Logistic Regression, SVM, Random Forest) and deep learning models (LSTM, BERT) using TF-IDF and Word2Vec.
 - Deployed a Flask web app integrated with LDA topic modeling to identify and visualize trending user sentiment themes.