NEERAJ GHATE

(720)-226-3190 | neerajrghate@gmail.com | linkedin.com/in/neeraj-ghate | github.com/neerajghate | neerajghate.vercel.app

Education

M.S. Computer Science, University of Colorado Denver (CGPA: 3.7) Expected May 2025

- **Coursework:** Algorithms, Operating Systems, Computer Architecture, Computer Vision, Deep Learning, Computer Networks, Natural Language Processing, Generative AI, Software Project Management, Data Science

B.E Electronics and Telecommunications, Pune University Aug 2018 - May 2022

Skills and Certifications

Programming Languages: C, C++, C#, Python, Java, JavaScript / TypeScript, HTML / CSS
Cloud & DevOps: Microsoft Azure (Storage, AKS, Functions, HDInsight, Cosmos DB), AWS (S3, EC2, Lambda, IAM), Kubernetes, Docker, Terraform, Jenkins, CI/CD, Public Cloud, Scalability, Distributed Systems, High-Performance Computing
Data & ML: Apache Spark, Hadoop, Apache Kafka, Snowflake, DynamoDB, SQL / NoSQL, ETL, Data Engineering, Data Analytics, Machine Learning, Deep Learning, Large Language Models (LLMs), TensorFlow, PyTorch, Keras, PowerBI, Tableau
Frameworks & Libraries: Node.js, React.js, Flask, Django, REST API, OpenShift Operators
Tools & OS: Linux, Git, Bash, VS Code, Terraform, Jenkins

Certification: AWS Certified Solutions Architect – Associate

Experience

Software Engineer, RealThingks GmbH July 2022 – July 2023

- Architected distributed training pipelines on Azure Kubernetes Service (AKS) and optimized inference latency by 30% for production LLM workloads.
- Built and optimized deep learning models using PyTorch and TensorFlow, contributing to efficient, production-ready pipelines.
- Led CI/CD and DevOps integration with Docker, Kubernetes, and Terraform, raising release frequency by 25%.
- Integrated Azure Blob Storage and Apache Kafka for scalable data ingestion supporting 100k images per minute.
- Conducted code reviews to fix critical C++ modules, improving code quality by 40% and overall performance.
- Delivered technical presentations to a 50-person team on scalability and HPC best practices.

Data Engineer Intern, Clairvoyant Aug 2021 – Sept 2021

- Migrated 5 TB on-prem Hadoop workloads to Microsoft Azure HDInsight, improving scalability and reducing cost by 40%.
- Optimized Spark jobs and designed OpenShift-based microservices to enable low-latency data access.
- Built and optimized data pipelines in Python and Scala, ensuring efficient data retrieval and 20% performance gains.
- Implemented data quality monitoring scripts with Kafka streams and Grafana dashboards.

Web Development Intern, Edify Accelerators Jun 2020 – Oct 2020

- Designed a user-friendly web application using ReactJS and deployed on Azure App Service, improving performance by 40%.
- Integrated frontend with Firebase for real-time updates and leveraged RESTful APIs hosted on Azure Functions.
- Created a proof-of-concept website showcasing key services using modern JavaScript frameworks and responsive layouts.
- Collaborated in agile sprints, demonstrating effective communication and flexibility across multi-disciplinary teams.

Projects

Personalized Recommendation System (LLM-Enhanced) Mar 2025 – Present

- Designed and developed a scalable recommendation engine for Netflix/Amazon-scale catalogs using collaborative filtering, content-based filtering, and deep learning models.
- Leveraged Azure OpenAI Service (GPT-4) for large language model embeddings, boosting cold-start precision by 18%.
- Implemented feedback loops and data analytics dashboards in PowerBI to surface user behavior insights to product teams.

Cloud-Enabled Full Stack Application Jan 2025 – Present

- Developed a microservices web application with React, Node.js, and Express, deployed via Docker and Kubernetes on Azure Kubernetes Service (AKS).
- Implemented autoscaling storage layer with Azure Cosmos DB and integrated CI/CD using GitHub Actions and Terraform.
- Instrumented distributed tracing and observability with OpenTelemetry, reducing MTTR by 35%.

Text Summarization with BART Aug 2024 – Dec 2024

- Built a hybrid text summarization pipeline combining TextRank and BART, fine-tuned on Python documentation to improve ROUGE scores by 25%.
- Containerized inference service on OpenShift and exposed REST endpoints for downstream analytics applications.