Vinay Pandya

San Francisco, CA|vinayharshadpandya27@gmail.com| (408)963-9802 | Linkedin | Google Scholar

INTRODUCTION

Software Engineer with 3+ years of experience building scalable web applications and machine learning solutions. Proven expertise in developing end-to-end systems using Java, Python, and modern web frameworks, with strong capabilities in cloud deployment, data engineering, and ML model implementation. Experienced in agile development practices and delivering user-focused products from concept to production

EDUCATION

Masters in Computer Science - San Jose State University, San Jose CA. GPA 3.6/4.0 May 2022

TECHNICAL SKILLS

Languages : Python, Java, C++, Node.js, R, Typescript.
Databases: Mysql, SQL, MongoDB, Athena, DynamoDB
Machine Learning: PyTorch, mlx-graphs, mlx **MLOPS:** Weights and Biases (Wandb), Weave, kubeflow
GPU Programming: Cuda, Metal(IOS)
Cloud: Azure, AWS
FrontEnd Frameworks: Flask, **React**, **FastAPI**, **Streamlit**.
Backend Frameworks: Python, PHP, **Node**, **Java**, **Kotlin**
Version Control: Git, Github
DevOps CI/CD: Docker, Kubernetes, **Github actions, Jenkins**, ARGOCD

PROFESSIONAL EXPERIENCE

Member of technical staff - Illumio **Apr 2025 - present**

- Created and deployed API gateway using spring boot to integrate multiple downstream services (15+)
- Used Horizontal Pod Autoscaling and efficient connection management to keep the system highly available (100k requests per second)
- Skills used: **Java, spring boot, HPA, kubernetes, Jenkins, argocd, API gateway, Azure, AWS**

Software Development Engineer - Amazon **Aug 2022 - Mar 2025**

- Deployed ML pipelines with NLP and Transformers for Alexa on **SageMaker**, improving data transcription automation from 15% to 20%.
- Built GDPR-compliant data pipelines using **AWS Glue** and SQL for Amazon's **A/B testing platforms**.
- Engineered scalable model serving infrastructure using **AWS Step Functions, DynamoDB, and Lambda**, processing 100K+ daily requests with 99.9% uptime.
- Built **SageMaker Studio extension** integrating **Hugging Face** and Papers with Code, reducing dataset access time from 90 to 2 days through automated validation and caching.
- Implemented **RAG** system using **AWS Bedrock and vector databases** for Amazon Payments onboarding, reducing onboarding time by 30%.

Associate Software Engineer - Accenture Services Private Limited: **Sep 2018 - Sep 2019**

- Built incident management chatbot with Node.js and Firebase, integrating PagerDuty for automated escalation and reducing incident response times by 20%.
- Created monitoring dashboards using Splunk and AWS CloudWatch to track business KPIs, improving financial visibility by 10%.

PROJECTS:

Mlx-cluster (Open source Software OSS):

- Created a **Python library** for generating and calculating random graphs efficiently on **Apple GPU** using **mlx-graphs and mlx**
- Utilized **metal kernels (cuda)** with Python for efficiently calculating random walks and biased random walks achieving **1.2x** performance compared to regular PyTorch kernels on Mac OS
- Used **Nanobind, Poetry and Python Packaging Index (PIP)** for publishing the python package.

Malware detection from Opcode Sequences and Bigrams (Masters Thesis)

- Created a **malware detection algorithm** which can detect a malware by disassembling opcode sequences.
- Utilized **BERT and XLNet** architectures and various other ensemble methods to improve the accuracy from **95% to 97%**.

Identifying Illicit transactions In Elliptic bitcoin network

- Utilized **PyTorch** and **Torch Geometric** GNN(Graph neural network) paradigms to identify illicit nodes in a blockchain transaction graph achieving an accuracy of **96%**.
- Used **Spectral graph theory, Graph Neural networks** and **cosmograph** to Identify illicit nodes.
- Utilized **Sklearn, Pandas and Matplotlib** to preprocess and visualize node features for transactions.
- Utilized **GNN explainer** to create explanations for machine learning predictions.