# SUDIPTA PATHAK

📞 +1 (848)-373-5136 ✉ sudipto.pathak@gmail.com 🔗 https://www.linkedin.com/in/sudipta-pathak-a6910117/ ⭘
https://github.com/sudiptap

## PROFESSIONAL SUMMARY

Senior Machine Learning Infrastructure Engineer with 8+ years of experience building and scaling ML platforms and distributed systems in production. Skilled in end-to-end ML lifecycle management, from feature stores and data pipelines to distributed training, deployment, and monitoring. Proficient in Kubernetes, Docker, ML orchestration (Kubeflow, Ray, Airflow), infrastructure-as-code (Terraform), GPU optimization, and Linux internals. Experienced in delivering secure, high-performance ML services in multi-region cloud environments while mentoring peers and collaborating across teams.

## TECHNICAL SKILLS

**Languages**: Python, C++, Java, SQL
**ML/DL Frameworks**: PyTorch, vLLM, Scikit-learn
**MLOps & Cloud**: AWS (SageMaker, EC2, S3, Lambda, ECS), MLflow, Kubeflow, Docker, Kubernetes
**Data Engineering**: Apache Spark, Kafka, Hadoop, Snowflake, Databricks, Apache Beam
**LLM & NLP**: LangChain, LangGraph, Hugging Face, OpenAI API, BERT, GPT, Transformers
**DevOps & Infrastructure**: Terraform, Jenkins, GitHub Actions, Prometheus, Grafana, Linux

## EXPERIENCE

**Senior Machine Learning Engineer** — Sept 2023 – Present
*JPMorgan Chase* — *Jersey City, NJ*

- Co-developed LLMSuite, an internal marketplace for deploying ML products into AWS, replacing ad-hoc, downtime-prone deployment processes with a scalable, standardized platform
- Led a team of 3 engineers to design and build a Natural Language to Query/Python code generation engine, enabling analysts and developers to interact with data and APIs without manual coding
- Scaled the system to support 300K+ internal users, ensuring high availability, automated scaling, and secure operation in a cloud-native AWS environment
- Engineered the platform's cloud infrastructure using Terraform and AWS services (ECS, Lambda, S3, API Gateway, CloudWatch), integrating agentic ML workflows built with LangChain, LangGraph, and OpenAI APIs

**Software Engineer II, Public Cloud Infrastructure** — Sept 2022 – July 2023
*Amazon Web Services* — *Jersey City, NJ*

- Led a team of 6 engineers to deliver AWS Glue support for large instance types (released in May 2023), enabling customers to run high-memory, high-throughput ETL workloads
- Designed and implemented a backend feature to resolve hot partition bottlenecks, improving service scalability and performance for distributed data workloads
- Eliminated a recurring customer pain point by introducing automated cleanup for leaked Elastic Network Interfaces (ENIs), reducing operational overhead and AWS resource waste
- Independently architected and drove the development of a latency-optimization feature that reduced Glue job startup times, improving time-to-insight for customers
- Participated in 24/7 on-call rotations, triaging and resolving high-priority, customer-facing production issues to maintain service reliability and SLAs

**Senior Software Engineer, Bloomberg Cloud Infrastructure** — July 2020 – Sep 2022
*Bloomberg* — *Princeton, NJ*

- Designed and implemented end-to-end pipelines and microservices for ingestion and parsing of large-scale financial and news data, enabling faster and more reliable downstream analytics
- Led the migration of critical services from legacy C++ architecture to an event-driven, asynchronous Python microservices framework backed by distributed storage
- Drove the containerization and deployment of backend services into Kubernetes, standardizing deployment workflows and improving service resilience and scalability
- Built and maintained production-ready, containerized services with automated monitoring, logging, and alerting for rapid issue detection and resolution

**Machine Learning Engineer** — Oct 2017 – July 2020
*Siemens Corporation* — *Princeton, NJ*

- Led a DARPA-funded research initiative to design and deliver scalable platforms for information extraction and document understanding in defense and intelligence applications.

- Architected and implemented end-to-end systems for advanced machine learning workflows, including processing and analysis of complex data types such as point cloud datasets.
- Directed cross-functional engineering efforts, ensuring timely delivery of research prototypes and alignment with evolving technical requirements.

## EDUCATION

**University of Connecticut**                                                                      Storrs, CT
*PhD in Computer Science*                                                                        *2011 – 2017*
- Focus: Artificial Intelligence and Machine Learning
- Dissertation: "Scalable Machine Learning Infrastructure for Real-time Applications"