Zahin Roja

zahinroja.us | zahinroja1@gmail.com | 929-476-8241 | New York, NY

SUMMARY

Data Engineer with 6+ years of experience designing, building, and deploying distributed data systems, real-time ingestion platforms, and analytics pipelines across financial services and technology. Hands-on experience with Python, SQL, Kafka, Spark, Kubernetes, and relational and analytical databases. Proven ability to own data pipelines through production delivery, develop containerized services, and partner with engineering and business stakeholders to deliver reliable data products.

RELEVANT WORK EXPERIENCE

Morgan Stanley, *Data Engineer,* **New York, NY** Jun. 2026 – Present

• Own end-to-end delivery of real-time Kafka pipelines that stream equity, options, and crypto trade data in parallel into SQL Server and SingleStore, supporting **100+** E*TRADE fraud flows and **hundreds** of fraud rules at **3-4 million** events daily.
• Develop a Python service deployed on Kubernetes that processes and persists raw payment transactions in real time, giving downstream consumers the data they need to detect fraud.
• Integrate SingleStore data with downstream analytics workflows and troubleshoot ingestion, payload, schema, persistence, and data-availability issues across QA and production environments.
• Lead sprint initiatives within a 5-person real-time data engineering squad, translating technical work into stories, coordinating dependencies, and driving cross-team follow-through.

Bank of America, *Data Engineer,* **New York, NY** Jul. 2020 – Jun. 2026

• Built and deployed containerized services (Python/Django) on Kubernetes/OpenShift to expose platform data via APIs, improving credit underwriting decision rates from **25% to 75%** globally.
• Designed and maintained large-scale data ingestion and transformation pipelines using Spark, SQL, and ETL frameworks, optimizing partitioning and indexing to reduce reporting latency by **20%** across enterprise systems.
• Migrated and tuned workloads between Photon and Spark compute engines, achieving **20% faster data processing** for 30,000+ client-facing transactions daily.
• Created automated monitoring and alerting for batch workflows using Splunk and Autosys, ensuring SLA compliance during 3–4 monthly production releases.
• Collaborated with data science teams to apply statistical analysis in engine performance comparisons, influencing adoption strategies.
• Mentored two interns on scalable data processing best practices, resulting in both being hired full-time.

ViacomCBS, *Data Science Intern,* **New York, NY** Jan. 2020 – Mar. 2020

• Created 15+ PySpark ETL scripts centralizing real-time social media data for campaign analysis.
• Built/trained AWS SageMaker models, improving ad targeting and campaign effectiveness by **10%**.

SKILLS, CERTIFICATIONS, AND LANGUAGES

Big Data, Streaming & Distributed Systems: Apache Kafka, Apache Spark, PySpark, Databricks (Photon), ETL/ELT Pipelines
Programming & Frameworks: Python, SQL, C++, Django, REST APIs
Databases: MSSQL, SingleStore, DynamoDB, MongoDB, MySQL, Oracle
CI/CD, Monitoring & Scheduling: Jenkins (CI/CD), Splunk, Autosys, Grafana, OpenTelemetry
Cloud and Infrastructure: AWS, Azure, GCP, Kubernetes, Docker, OpenShift
Certifications: Generative AI with Large Language Models, DeepLearning.AI and AWS (Feb 2024)
Languages: Native English, fluent Bengali, intermediate Spanish, conversational Hindi
Other Tools: JIRA, Git

EDUCATION

University of Southern California Online, *Master of Science in Data Science* May 2024
Fordham University New York, NY, *Bachelor of Science in Computer Science* May 2020