SWASTIK CHOWDHURY

swastiksc1996prof@gmail.com | 412-844-1548 | https://www.linkedin.com/in/swastik-chowdhury/

EDUCATION

Carnegie Mellon University **August 2024 - August 2025**
Master of Information Systems Management
Coursework: Machine Learning, Deep Learning, Machine Learning in Production, Database Systems, Distributed Systems, DevSecOps, NoSQL Databases

Maulana Abul Kalam Azad University of Technology **June 2015 - June 2019**
Bachelor of Technology
Coursework: Computer Programming, Data Structures & C, Object-Oriented Programming, Database Management Systems

WORK EXPERIENCE

Carnegie Mellon University, *Graduate Teaching Assistant* **January 2025 - May 2025**
Courses: Machine Learning in Production/AI Engineering, and Advanced Python

- Delivered lab sessions for 30+ students on end-to-end ML model deployment, monitoring, and MLOps best practices using Jenkins, Docker, Kubernetes, MLflow, Prometheus, and Grafana.
- Guided 4 student teams (20 members) in building scalable recommendation systems, mentoring them on data drift detection, Responsible AI (fairness, explainability, security), deployment experimentation, and ML architecture design.
- Enabled 20+ working professionals to advance their Python skills by providing hands-on guidance in data preprocessing, exploratory data analysis, API integration, and relational database management.

Vertisage Technologies, *Senior Data Engineer* **April 2023 - March 2024**

- Led a team of 5 to migrate Redshift procedures to Spark on EMR as part of a cost optimization initiative, reducing annual expenses by over $1 million for a US-based ad tech company.
- Designed and implemented a real-time Change Data Capture (CDC) pipeline using Debezium, Apache Kafka, and Apache Iceberg, eliminating full-table scans and improving data freshness from daily batches to near real-time.
- Built an event-driven, serverless failure alerting pipeline with Amazon SNS, SQS, and AWS Lambda, enabling real-time email and Slack notifications for production incidents.

Tata Consultancy Services, *Data Engineer* **July 2019 - April 2023**

- Optimized Spark Scala jobs for an automotive company's connected vehicle telematics data, using broadcasts, caching, and efficient joins to reduce cloud compute costs by 15% for real-time fleet insights.
- Implemented Spark ETL pipelines orchestrated with Apache Airflow, ingesting 5GB+ of vehicle telemetry into AWS Athena, replacing manual data loads and saving over 10+ hours of effort per week.
- Owned around-the-clock production support with PagerDuty, diagnosing and resolving pipeline incidents under tight SLAs to safeguard 100% data integrity for real-time analytics.
- Enforced strong data quality and testing standards with unit tests and Deequ, cutting production data errors and preventing regressions across ETL pipelines.

ACADEMIC PROJECTS

Machine Learning in Production for RecSys

- Designed and deployed a scalable movie recommendation system (1M users, 27K movies, <500 ms latency, 99% availability) using KNN, RF, and TF-IDF models evaluated on Precision@K, Recall@K, MRR, hit rate, and latency.
- Built real-time data pipelines with Kafka and Airflow, developed a feature store with Feast, and ran A/B tests to optimize model performance.
- Containerized Flask services with Docker, deployed them to Kubernetes using Jenkins CI/CD, and ensured monitoring (Grafana, Prometheus) and ML traceability with MLflow.

Healthcare Fraud Detection System

- Built an async ELT pipeline with checkpointing to fetch paginated healthcare fraud data (~1M prescribers) yearly, using a medallion architecture to clean, transform, and curate data with high integrity.
- Applied time-series imputation, feature scaling, SMOTE, and temporal validation to train fraud detection models (Random Forest AUC 0.65, Accuracy 0.79) tackling extreme class imbalance in a $100B US fraud landscape.

Technical Blog

- Popular articles include Modernizing Data Ingestion for Machine Learning , The Art of Troubleshooting Data Pipelines.

SKILLS

Programming Languages	Python, Scala, Java, SQL
Big Data Frameworks	Apache Spark (Core, SQL, Streaming), Hadoop, Kafka
Databases & Data Warehousing	Redshift, BigQuery, Iceberg, Snowflake, Hive, Trino, Cassandra, MongoDB
Data Modeling & ETL Tools	Dimensional Modeling (Star/Snowflake), dbt, Apache Airflow, AWS Glue
DevOps & Cloud Platforms	AWS, GCP, Docker, Kubernetes, Jenkins, Linux, Bash, Prometheus, Grafana, Git