SWASTIK CHOWDHURY
swastiksc1996prof@gmail.com | 412-844-1548 | Linkedin | Medium | Github | Website

EDUCATION

Carnegie Mellon University, Heinz College, Pittsburgh, Pennsylvania — August 2025
Master of Information Systems Management
Machine Learning, Deep Learning, Unstructured Data Analytics, Business Analytics, Applied Econometrics, Statistics
Machine Learning in Production, Database Systems, Distributed Systems, DevSecOps, NoSQL

Maulana Abul Kalam Azad University of Technology, Kolkata, India — June 2019
Bachelor of Technology

EXPERIENCE

Carnegie Mellon University, Pittsburgh, Pennsylvania — January 2025 - May 2025
Graduate Teaching Assistant, Machine Learning in Production/AI Engineering

- Mentored teams on data drift, Responsible AI (safety, fairness, transparency), and scaling ML systems using different architectures.
- Led recitations on Jenkins, Docker, Kubernetes, Kafka, Grafana, Prometheus, and MLflow, for model deployment and monitoring.
- Demonstrated model explainability using SHAP and LIME, and led safe experimentation via A/B testing and canary deployments.

Carnegie Mellon University, Pittsburgh, Pennsylvania — January 2025 - March 2025
Graduate Teaching Assistant, Advanced Python

- Assisted students with data preprocessing using pandas and NumPy, retrieving data via web APIs, and managing relational databases

Vertisage Technologies, Bangalore, India — April 2023 - March 2024
Senior Data Engineer

- Led a migration of **Redshift** procedures to **EMR** using Spark, resulting in over $1 million in annual cost savings.
- Engineered a tool for indexing documents with **LlamaIndex**, using **RAGs** for knowledge retrieval.
- Optimized **Spark Scala** job execution time by ~25% through ideal partition sizing and shuffle parameter tuning.
- Established external tables in **Snowflake**, integrating data stored in S3; devised upsert ability leveraging **Delta Lake**.

Tata Consultancy Services, Hyderabad, India — July 2019 - April 2023
Senior Data Engineer

- Designed a data ingestion pipeline with Debezium, **Apache Kafka**, and **Apache Iceberg** for efficient **Change Data Capture**.
- Enhanced Spark Scala jobs using broadcasts, caching, resource allocation, and optimized joins resulting in 15% cost reduction.
- Deployed infrastructure leveraging **EMR** on **EKS** and Airflow, enhancing system scalability and operational efficiency.
- Resolved critical data pipeline issues using **Postman** for API debugging and systemic root cause diagnosis for data integrity.

Data Engineer

- Built a secure Data lake using Amazon S3, with ETL orchestration via **AWS Glue** and automated ingestion via **Lambda** triggers.
- Implemented ETL pipelines for Spark jobs leveraging **Apache Airflow** to load 5GB+ of Parquet data into **AWS Athena**.
- Ensured data quality and code coverage through SparkTestingBase unit testing, Deequ, and data cleaning techniques.
- Integrated **Tableau** dashboards with live data sources via REST API enabling KPI tracking and data-driven business decisions.

ACADEMIC PROJECTS

Machine Learning in Production for RecSys — August 2024 - December 2024

- Designed a scalable system for **1M** users and **27K** movies, with **500** ms latency and **99**% availability.
- Evaluated KNN, RF, and TF-IDF + Cosine Similarity for recsys based on Precision@K, Recall@K, MRR, hit rate, and latency.
- Leveraged **Kafka** for real-time data ingestion and streaming at scale; implemented data pipelines using **Airflow**.
- Developed a **feature store** using feast, and ran **A/B tests** to evaluate and improve model performance.
- Containerized Flask services with **Docker** and built CI/CD pipelines to deploy them on **Kubernetes**.
- Managed deployments using **Jenkins**, with monitoring via **Grafana** and **Prometheus** and ensured ML traceability with **MLflow**.
- Promoted fairness via data checks, bias detection, and metrics like Equal Opportunity and exposure balancing.
- Ensured system security and fairness via threat modeling (**STRIDE**), rate limiting, anomaly detection.

Healthcare Fraud Detection System — August 2024 - December 2024

- Built an async ELT data pipeline with checkpointing to fetch paginated healthcare fraud data reliably by year and category.
- Designed a **medallion** architecture to clean, transform, and curate fraud datasets with strong data integrity.
- Applied time-series imputation, feature scaling, and temporal validation to train and compare ML models using **SMOTE**.
- Used **BigQuery** for data warehousing and **Looker** for fraud analytics dashboards.

Data Intensive Scalable Microservices for E-commerce — March 2025 - April 2025

- Implemented a microservices-based AWS bookstore app using Node.js on **EKS** using the **CQRS** pattern.
- Provisioned infrastructure with CloudFormation (subnets, EC2, ALBs).
- Optimized data storage with **MongoDB** for fast reads (300ms) and **MySQL RDS** for writes, using the **BFF** pattern.
- Improved system resilience via **circuit breakers** and Kafka-based event workflows, with automated CRM email integration.

SKILLS

Programming Languages :	Python, Scala, Java, SQL
Big Data & ML Frameworks :	Spark, Hadoop, Kafka, Machine Learning, Deep Learning, NLP, Gen AI
Databases & Data Tools :	Oracle SQL, MongoDB, Redshift, BigQuery, Iceberg, Snowflake, Hive, Trino, Cassandra, Redis, Neo4j
DevOps & Cloud:	AWS, GCP, Docker, Kubernetes, Jenkins, ArgoCD, Ansible, Istio, Linux, Bash, Prometheus, Grafana
Data Engineering Tools:	Airflow, dbt, MLflow, Postman, Tableau, Looker, Git