Sourav Halder

Data Scientist

+91-8777893442 | halder.sourav1996@gmail.com | linkedin.com/in/sourav–halder | github.com/souravhalder1996 | souravhalder.dev

Summary

Data Engineer with 4+ years of experience designing scalable data pipelines and cloud-native solutions, now applying advanced analytics and machine learning to solve business problems. Holds a Master's degree in AI and Robotics, with strong proficiency in Python, SQL, AWS, and Machine Learning. Skilled in transforming complex, multi-source data into model-ready assets, building predictive solutions, and driving strategic decisions through experimentation and statistical analysis. Proven ability to collaborate across engineering, product, and business teams to deliver measurable impact in fast-paced environments.

Skills

Programming & Data Engineering: C, Python, SQL, Databricks, ETL/ELT Pipelines, Git
Machine Learning & AI: NumPy, Pandas, Scikit-learn, TensorFlow, Keras, XGBoost, MLflow, Feature Engineering, Model Deployment, n8n, Portkey AI, Hugging Face, LangChain
Cloud & Deployment: AWS (S3, EC2, ECR, Lambda, Glue, Redshift, SageMaker, Bedrock, CloudWatch), Docker, CI/CD
Analytics & Visualization: Matplotlib, Seaborn, SQL Analytics, Statistics

Experience

Infosys Limited **October 2021 – Present**
Senior Associate Consultant

- Engineered and maintained 50+ production ETL pipelines using Python, SQL, and AWS Glue, orchestrating data extraction from multi-source SAP ERP, SQL Server, and REST APIs into Redshift data warehouse supporting 14 data marts, ensuring timely data availability for business dashboards and analytics teams.
- Implemented robust pipeline orchestration with event-driven and schedule-based triggers (Lambda, EventBridge, S3), incorporating programmatic retry logic and data validation checks that eliminated synchronization failures by 85%, resolved data mismatches, and achieved 99.5% pipeline reliability.
- Optimized data delivery cycles to support monthly Sales & Operations Planning (S&OP) processes for leadership and regional sales heads, cutting data delays by 60% and ensuring accurate, timely insights for strategic sales planning and forecasting.
- Re-architected data extraction pipeline by migrating from direct SQL database to Apache Iceberg-based live layer with incremental delta loads, achieving 93.75% reduction in data extraction and eliminating performance bottlenecks caused by high-volume ERP replication traffic.
- Streamlined incident management by integrating ServiceNow API for pipeline monitoring, enabling automatic ticket creation and intelligent team assignment for job failures; decreased mean time to resolution (MTTR) by 40% and eliminated manual triaging overhead.
- Architected a Generative AI PaaS integrating Portkey AI LLM Gateway (routing, rate limiting, serving, usage tracking) and AWS Bedrock for centralized RAG capabilities; secured multi-tenant access via Active Directory (AD) and SailPoint to allow teams to securely ingest and query proprietary data.
- Implemented custom n8n workflows, engineering a Jira ticketing agent that queries Bedrock Knowledge Bases to auto-suggest resolutions; designed LLM-generated stakeholder reports and a sync workflow from SharePoint to Confluence/Web.

Education

Jadavpur University **August 2018 – July 2021**
M.Tech in Intelligent Automation and Robotics *Grade: 90.36%*

Brainware Group of Institutions **August 2013 – July 2017**
B.Tech in Electrical Engineering *CGPA: 8.49/10*

Projects

Network Security System | ***Python, XGBoost, Optuna, MLflow, AWS, Docker*** GitHub | Live View

- Architected an end-to-end network threat detection pipeline using XGBoost, Random Forest, and Stacking Classifiers; leveraged SMOTE and Optuna to achieve 97.9% Test Recall with <15ms inference latency.
- Deployed containerized (Docker) models on AWS (EC2/S3/ECR) using CloudFormation IaC and CI/CD; integrated MLflow for Champion-Challenger version promotion and Evidently AI for data drift monitoring.

Kidney Disease Classification | ***Python, Deep Learning, MLOps, Flask, MLflow, DVC*** GitHub | Live View

- Developed an EfficientNetV2-B0 classifier for kidney tumors on 7.3K+ CT scans using TensorFlow, achieving 100% validation accuracy (1.00 AUC) via a 2-stage transfer learning strategy and label smoothing.
- Orchestrated a robust MLOps pipeline using DVC and MLflow/DagsHub for reproducible tracking, deploying the containerized API via Docker and GitHub Actions on AWS EC2.

Awards & Certifications

- **AWS Certified Solutions Architect – Associate**: View Credential
- **AWS Certified Developer – Associate**: View Credential
- **Databricks Certified Data Engineer Professional**: View Credential
- **Databricks Certified Machine Learning Engineer Associate**: View Credential