

SHREELA SAPKOTA

Kathmandu, Nepal | sapkotashreela02@gmail.com
linkedin.com/in/shreelasapkota | github.com/shreelasapkota

SUMMARY

Associate Data Engineer with hands-on experience building ETL pipelines and data-quality frameworks for large-scale US healthcare datasets using **PySpark, SQL, and AWS (S3, Athena)**. Skilled in HIPAA-compliant processing of healthcare pricing, claims, and provider data, from raw ingestion to validated, analytics-ready datasets.

EXPERIENCE

Associate Data Engineer

Deerhold Ltd.

April 2025 – Present

Kathmandu, Nepal

- Automated **40+ CMS healthcare reference-data ETL pipelines** (ICD-10, CPT/HCPCS, MS-DRG, NPI, fee-schedule crosswalks) in a config-driven **Python/PySpark** framework, cutting the full multi-million-record refresh cycle from **24+ hours to ~2 hours — a 12× speedup**.
- Built **end-to-end pipeline automation** for provider NPI, IPF-PPS, and hospital-quality datasets: **Selenium**- and config-driven source acquisition from CMS portals, **S3** ingestion, PySpark transformation, and publication to **Athena**-queryable tables.
- Authored **automated data-quality test suites** running SQL validations on **AWS Athena** (schema, mapping, and prod-vs-UAT comparison checks) with formatted Excel reports auto-published to Google Drive, replacing manual QA of million-row master tables.
- Researched **drill-down logic** across existing healthcare sources (CMS fee schedules, provider registries) to derive new **analytics-ready datasets** — evaluating source structures, defining mapping rules, and validating each derived table before release.
- Partner with data engineering and compliance teams to enforce **HIPAA-compliant** data handling across all client-facing healthcare datasets.

Survey and Data Collection Team Member

Kathmandu Metropolitan City Project

September 2024 – October 2024

Kathmandu, Nepal

- Collected images of inscriptions from religious sites to aid in language translation and preserve ancient history.
- Gained early insight into **data collection methodology** for machine learning, reinforcing the importance of comprehensive, high-quality data for effective model training.

PROJECTS

Heart Disease & Diabetes Prediction | Python, Scikit-learn, Pandas

- Built predictive models (**Logistic Regression, Random Forest, KNN, Neural Networks**) for disease-risk assessment; applied feature engineering, hyperparameter tuning, and ensemble methods.
- Evaluated models with **ROC-AUC** and confusion matrices to select the best-performing classifier.

Lyrics Genre Detection | Python, Scikit-learn, NLP

- Developed an ML pipeline predicting song genres from lyrics: text preprocessing, **TF-IDF vectorization**, and a Random Forest classifier evaluated via accuracy and classification reports.

SQL Sales Analytics | PostgreSQL, Python

- Modeled a **1.55M-row** e-commerce dataset (9 tables) with a Python ETL loader; wrote **window function**, CTE, and join queries for ranking and trend analysis, cutting runtime **3–11×** via **EXPLAIN ANALYZE** profiling and indexing.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript/TypeScript

Data Engineering: PySpark, ETL Pipelines, AWS (S3, Athena), Trino/Presto, Boto3, Data QA & Validation, Web Scraping

Databases: PostgreSQL, MySQL, MongoDB

Data & ML Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Tools: Git, GitHub, DBeaver, Postman

Domain: HIPAA-compliant healthcare data (pricing, claims, provider identifiers)

EDUCATION

Kathmandu University

Bachelor of Engineering in Computer Engineering, CGPA: 3.67/4.00

Feb 2021 – Aug 2025

Dhulikhel, Nepal

AWARDS & COMMUNITY

2nd Runner-Up, Hult Prize at Kathmandu University – pitched Myra, a women's health-tracking app.

British Airways Data Science Job Simulation (Forage) – customer-review analysis and predictive modeling.

Designer, Women in Data Science (WiDS) Kathmandu – event promotion and community outreach.