Suyog Mainali

Williamsburg, VA · +1 (757) 206-8339 · smainali01@wm.edu · linkedin.com/in/suyog41 · Portfolio · GitHub

PROFESSIONAL SUMMARY

Data Engineer (MSBA '26, William & Mary) with 5+ years building large-scale ETL/ELT systems, including pipelines that process 100TB+ of healthcare data daily across distributed PySpark, SQL, and AWS environments. Experienced in data lakes, warehouse modeling (star/snowflake), and event-driven ETL using Kafka, Lambda, and Step Functions. Skilled in Airflow, dbt, Docker/Kubernetes, and CI/CD, with a strong focus on data quality, lineage, and production-grade analytics for ML and BI applications.

EDUCATION

William & Mary, Raymond A. Mason School of Business, Williamsburg, VA, USA August 2025 - May 2026
Master of Science, Business Analytics (**Expected May 2026**)

Vellore Institute of Technology (VIT), Vellore, India August 2018 - May 2022
Bachelors of Technology, Computer Science Engineering

PROFESSIONAL EXPERIENCE

Zakipoint Health, Sanepa, Lalitpur, Nepal July 2020 - August 2025
Senior Data Engineer

- Built ETL/ELT pipelines in Python, PySpark, and SQL to process 100+ TB of multi-source healthcare data (claims, eligibility, EHR) daily into AWS S3 data lakes, enabling analytics, ML, and reporting across teams.
- Deployed dimensional (star/snowflake) models to Redshift/Snowflake, tuning distribution/sort keys and clustering for performance optimization.
- Developed data quality, monitoring, and SLA tracking frameworks for near-real-time BI updates with 99.8% uptime.
- Automated deployments using Git, Jenkins CI/CD, Docker, and Kubernetes, enabling reliable production rollouts and environment-based promotions.
- Built provider cost-estimation and search workflows using PostgreSQL + Elasticsearch for fast retrieval of claims, provider prices, deductibles, and OOP calculations.
- Developed Airflow DAGs and dbt models (incremental, tested, documented) in an Agile/Scrum environment, delivering iterative data platform features across teams.

Om Hospital, Kathmandu, Nepal April 2018 - July 2020
Data Engineer

- Built data warehouse and data lake pipelines with Python, SQL, and Azure ADLS, enabling data consolidation across clinical and operational systems.
- Developed data marts for executive dashboards in Tableau, improving operational decision-making and reducing patient wait times by 25%.
- Standardized schemas, indexing, and partitioning; improved query performance by 30%.
- Implemented data governance documentation, data dictionaries, and validation frameworks.

CORE SKILLS & COMPETENCIES

- Languages: Python, SQL, PySpark
- Data Engineering: ETL/ELT, Event-driven ETL (Kafka, Lambda, Step Functions), Data Pipelines, Data Lakes, Data Warehousing
- Orchestration: Airflow, dbt, Jenkins, CI/CD
- Modeling & Storage: Redshift, Snowflake, PostgreSQL, MySQL, MongoDB, Elasticsearch, Parquet/JSON
- Streaming: Kafka
- Cloud: AWS (S3, EC2, IAM, Glue, Lambda, EMR, Step Functions), Terraform, Azure ADLS
- DevOps: Docker, Kubernetes, Linux/UNIX, Shell Scripting (Bash), Monitoring, Logging, SLA Management
- Reporting: Tableau, REST APIs, Automated Reports

PROJECTS

- **Amazon Reviews ETL Pipeline (SerpApi + Playwright)** – Built a full-stack ETL pipeline automating extraction, validation, and synchronization of 500K+ Amazon product listings using SerpApi, Playwright, async scraping, retry/resume logic, and a Flask API service; stored normalized metadata and reviews in PostgreSQL.
- **Siftline – Grounded Document QA** – Developed a private, grounded Q&A system using FAISS + TF-IDF hybrid retrieval, Transformers-based summarization, and secure PDF/DOCX/TXT ingestion; deployed as a Streamlit app with Azure-ready, cloud-native architecture.
- **$100 Question – Dash Analytics App** – Created an interactive macro-economics dashboard using Plotly Dash; streamed financial indicators, rendered scenario-based charts, and deployed the application with CI/CD on Render.