Saharsh Kamod Koli

Fairfax VA,22030. | saharshkoli7007@gmail.com |skoli2@gmu.edu | (571)4283187|Linkedin

---

Professional Summary

Python-focused Data Engineer experienced in building data pipelines, ETL workflows, and ML-powered data systems using AWS and cloud-native technologies. Skilled in Python, SQL, LLM-based data processing, and structured/unstructured data modelling across analytics, NLP, and GenAI solutions.

Experience

Yes Tech Corp. — New York, NY (Remote)

Data Analytics Engineer | 07/2025 – Present

- Engineered AWS-based data pipelines using Python, DynamoDB, and Athena to ingest, transform, and query user interaction data from alpha and beta product launches.
- Designed automated data processing workflows to clean, aggregate, and analyse structured and semi-structured survey data, improving insight delivery latency by ~25%.
- Partnered with product and leadership teams to translate business requirements into scalable analytics and data models, identifying 5+ optimization opportunities.

Siemens EA Factory — Goa, India

Data & Reporting Intern (Database Management) | 06/2022 – 07/2022

- Designed and maintained a structured reporting database using relational data modelling concepts with MS Excel and Visual Basic, managing 800+ training records for operational tracking and analysis.
- Automated data entry and filtering processes, cutting manual reporting effort by 30% and improving data accuracy and accessibility.
- Developed custom dashboards, reports, and presentations used by 5+ managers to evaluate performance, monitor training outcomes, and support operational decision-making.

Numerouno Education Marketing Services Pvt. Ltd. — Mumbai, India

Business & Data Support Intern | 11/2022 – 01/2023

- Reviewed customer interaction data while assisting 50+ clients weekly, achieving 95% issue resolution within 24 hours and improving customer satisfaction by 15%.
- Supported performance tracking and weekly reporting across 50+ client interactions, identifying recurring trends and communicating insights through verbal and written communication to internal stakeholders.

Projects

Air Traffic Controller (ATC) Voice Analytics Dashboard | 07/2025 – 12/2025

- Designed and implemented a Python-based data pipeline using a custom ATLAS algorithm to process unstructured audio, generate structured datasets, and support LLM/NLP-driven safety and trend analysis.
- Built data ingestion, transformation, and enrichment layers for 100+ ATC audio samples, applying segmentation, noise reduction, and transcription to improve downstream analytics and model readiness.
- Worked as a Developer in an Agile framework, contributing to biweekly sprints, stand-ups, and iterative analytics validation with a cross-functional team of 5+ members.

Database Querying & Reasoning for Financial QA (Fin-DBQA) | 09/2024 – 11/2024

- Engineered a financial data querying and reasoning system using Python, DBQA, and NLP, enabling automated analysis of structured financial datasets.
- Extracted entities, measures, and constraints from natural-language queries using spaCy, improving data retrieval accuracy and analytical reasoning across 100+ test queries.
- Integrated Hugging Face Transformers to support LLM-based entity extraction and semantic reasoning, achieving 79% query accuracy across 100+ financial queries.

Bridge Condition Analysis & Predictive Maintenance | 10/2024 – 11/2024

- Examined large-scale infrastructure datasets (10k+ records) to identify key drivers of bridge condition and maintenance urgency, supporting data-driven resource allocation.
- Applied regression and statistical analysis to correlate weather, budget, and structural features with maintenance outcomes and cost forecasts, improving planning accuracy.
- Proposed data-driven maintenance strategies, resulting in a project paper submission to the FHWA InfoBridge competition.

Cybersecurity Threat & Vulnerability Analytics (EDA Project)
- Conducted exploratory data analysis on 1,000+ CVE-style vulnerability records using R (tidyverse, ggplot2, corrplot), performing data cleaning, preprocessing, and feature analysis.
- Assessed severity, exploitability, impact, and EPSS scores, generating correlation heatmaps and time-series trends (2015–present) to identify high-risk vulnerability patterns.
- Delivered risk prioritization insights using statistical summaries supporting data security and visual analytics, supporting vulnerability assessment and mitigation planning.

Technical Skills
- **Programming & Query Languages:** Python (data engineering, pipelines, automation), R (tidyverse), SQL (advanced querying, joins, aggregations)
- **Data Analysis & Engineering:** Data Pipelines (Batch & Streaming), Structured & Unstructured Data Processing, Data Cleaning & Wrangling, Hadoop, Spark, BigQuery, Apache Kafka, Data Lake Architecture
- **Cloud & Analytics Platforms:** AWS (S3, Athena, Redshift, Lambda), Google Cloud Platform (BigQuery), Microsoft Azure
- **Data Visualization & BI Tools:** Tableau, Power BI, Google Data Studio, Matplotlib, Seaborn, ggplot2
- **Databases:** Relational (PostgreSQL, MySQL (Oracle – familiar)), NoSQL (MongoDB)
- **Machine Learning & NLP (Applied):** LLMs / Generative AI (Hugging Face), spaCy, TensorFlow (foundational)
- **Tools & Version Control:** Git, Jupyter Notebook, VS Code

Education
George Mason University, Fairfax VA, United States (1/24 - 12/25)
- Master of science in Data analytical engineering. GPA - 3.90/4

University of Mumbai, India. (6/19 - 6/23)
- B.E. in Electronics and telecommunication Engineering. GPA - 3.46/4

Honors and Activities
- Attended AWS Summit Washington, D.C., gaining exposure to generative AI, database modernization, and cloud analytics, including Amazon Aurora, AWS DMS/SCT, Amazon Bedrock, and enterprise-scale deployments.
- Attended the AI + Robotics Summit in 2024 by SCPS in Washington, DC, to gain insights into cutting-edge technologies.
- Submitted a project paper on bridge maintenance and prediction to InfoBridge FHWA for their competition.