

Shreyas Shende

New York Metropolitan Area | shreyasshende222@gmail.com | (862)-409-0975 | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

Skills

- **Data Engineering & Cloud** – Airflow | dbt | Docker | Databricks | AWS (S3, Glue, Athena, Redshift) | BigQuery | Star Schema | Snowflake
- **Programming & SQL** – Python (Pandas, NumPy, PySpark) | SQL (MySQL, PostgreSQL, Window Functions, CTEs) | Matplotlib/Seaborn
- **Analytics & BI** – PowerBI (DAX) | Tableau | Excel/Sheets (Pivot tables, Power Query) | A/B Testing | Hypothesis & Regression testing
- **Machine Learning (Applied)** – Scikit-learn | Tensorflow | PyTorch | Time-Series Analysis | XGBoost
- **Collaboration & DevOps** – Git/GitHub | CI/CD | Linux | CRON | Confluence | JIRA | FastAPI/Flask

Experience

- Data Engineer | Morgan Stanley | New York, NY** Dec 2025 – Present
- Designed and automated production **ETL pipelines** integrating data from **7+ enterprise APIs** (Jira, Rally, OpenPages, etc.), processing **~35K records daily** through scheduled **AutoSys** workflows using **Python** and **SQL**, eliminating manual execution, reducing pipeline failures by **30%**, and improving data reliability, resiliency, and quality through validation, structured logging, and fault-tolerant error handling.
 - Built **Agentic AI** workflows leveraging enterprise **GPT** services to automate classification of **100+ monthly operational risk records**, identifying cloud providers, cloud services, business impacts, and AI-related issues using multiple enterprise knowledge sources while supporting human-in-the-loop validation for audit readiness.
 - Developed internal **AI agents** for Jira issue classification and automated **pull request reviews**, identifying security risks, coding standard violations, exposed secrets, and performance concerns before deployment, improving engineering quality and consistency across the team.
 - Automated **CI/CD** processes by configuring and maintaining **15 Jenkins pipelines**, deployment workflows, and Python dependency upgrades while building operational **Power BI** dashboards and engineering database fallback strategies over **2M+ records** to improve monitoring, reporting, and platform reliability.
- Graduate Research Assistant (Master's Project) | New Jersey Institute of Technology | Newark, NJ** Jan 2025 – May 2025
- Built an end-to-end RNA-seq analysis pipeline in Python and PyTorch Geometric, automating data preprocessing, graph construction, model training, and feature selection across Cervical, Kidney, Alzheimer, and Lung cancer datasets; reduced feature space by **>90%** and improved classification accuracy by **5–10%** over DESeq2 and EdgeR.
 - Led a 3-member research team in collaboration with Brown University's Alpert Medical School to source and validate clinical **RNA-seq datasets**, resulting in a **co-authored** research paper demonstrating the pipeline's adaptability across diverse genomic datasets.
- Platform Engineering & Automation Intern | Vendorpass (FIS Global) | Remote, US** May 2024 – Nov 2024
- Automated Python and Bash **ETL workflows**, reducing manual mapping effort by **90%** and improving **pipeline reliability**.
 - Built a **Streamlit dashboard** integrated with Oracle REST APIs to automate reporting across **20K+ records** while implementing CI/CD validation and post-clone automation that reduced environment **downtime from 4+ hours to 1 hour**.

Projects

- Ad Analytics Pipeline (Airflow + Databricks + BigQuery + Power BI) | [GitHub](#) |** Jul 2025 – Aug 2025
- Built a containerized **Python + Airflow pipeline** to ingest, validate, and orchestrate raw ad campaign data into **BigQuery**, ensuring scalable and automated cloud-native ETL.
 - Designed transformations in **Databricks (PySpark)** to clean and aggregate campaign KPIs (CTR, CPC, ROI) and delivered an interactive **Power BI dashboard** for real-time performance insights.
- Spotify Listener Analytics Dashboard (Power BI + Spark SQL) | [GitHub](#) |** May 2025 – Jun 2025
- Designed a star-schema data model by transforming nested JSON into Parquet using Python, enabling scalable analytics with **Spark SQL**.
 - Developed an interactive **Power BI** dashboard with DAX-based KPIs for engagement, churn, and revenue trends, reducing reporting turnaround by **70%**.
- Reddit Data Ingestion & Analytics Pipeline (AWS + Airflow) | [GitHub](#) |** Mar 2025 – Apr 2025
- Built a scalable ETL pipeline using **Airflow, Docker, Celery, and AWS** to ingest Reddit data via APIs into **Amazon S3** and orchestrate transformations with **AWS Glue and Athena**.
 - Automated ingestion of **100+ daily Reddit posts** into **Amazon Redshift**, enabling **SQL-based analytics** for trend, sentiment, and KPI reporting with minimal manual intervention.

Education

- M.S. Computer Science** | New Jersey Institute of Technology | GPA 3.95/4.0 Aug 2023 – May 2025
- B.E. Computer Engineering** | University of Pune | GPA 3.66/4.0 Aug 2019 – Jun 2023

Certifications & Publications

- AWS Certified Cloud Practitioner (CLF-C02) | Microsoft Certified: Azure Data Fundamentals (DP-900) | Oracle Cloud Infrastructure 2025 Generative AI Certified Professional**
- A Comprehensive Study on Simultaneous Localization and Mapping (SLAM): Types, Challenges, and Applications (ICSCSS 2023) (IEEE) | [Paper](#)**