SHASHANK BANALA

+1 984-286-8221| shashankbanala49@gmail.com | LinkedIn

SUMMARY

Data Engineer / Data Analyst with 4+ years of experience building cloud-based data pipelines, warehouse models, and ETL workflows across AWS, Azure, and Snowflake. Skilled in Python, SQL, and Spark for automating data ingestion, transformation, and data quality workflows. Experienced in applying machine learning and GenAI tools including OpenAI APIs and vector databases—to support anomaly detection, data validation, and reporting automation. Recognized for clear communication, analytical problem-solving, and developing data solutions that bridge engineering accuracy with business needs.

KEY ACHIEVEMENTS

- Migrated 12+ legacy pipelines to AWS, reducing server costs and simplifying maintenance.
- Optimized ETL and dashboard refreshes, helping finance teams close reports 2 days faster each month.
- Explored vector-based search (Pinecone, FAISS) to make querying unstructured data more intuitive during graduate research.

TECHNICAL SKILLS

Data Engineering & Processing: ETL Pipelines, Data Modeling, Apache Airflow, dbt, PySpark, Databricks, Delta Lake
Cloud Platforms: AWS (S3, Glue, Redshift, Lambda), Azure (Data Factory, Synapse, OpenAI Service), GCP (BigQuery, Vertex AI)
Databases & Warehousing: Snowflake, PostgreSQL, MySQL, SQL Server, MongoDB
Programming & Languages: Python, R, SQL, Bash
Data Science & Machine Learning: Regression, Classification, Clustering, Feature Engineering, Model Evaluation, A/B Testing, scikit-learn, Pandas, NumPy, Statsmodels
Generative AI & NLP Tools: OpenAI APIs, LangChain, Hugging Face, Vector Databases (Pinecone, FAISS), RAG Pipelines
Visualization & BI Tools: Tableau, Power BI, Excel (Pivot Tables, Macros), Power Automate
Orchestration & DevOps: Jenkins, Docker, Git, CI/CD Pipelines
Agile Collaboration: JIRA, Confluence, Scrum, Kanban

PROFESSIONAL EXPERIENCE

Data Engineer | *Inspira Financial, USA* **Oct 2024 – Present**

- Built and maintained real-time Databricks + Delta Lake pipelines to ingest and validate large volumes of financial records weekly, applying schema enforcement and time-travel to ensure data accuracy, consistency, and auditability across reporting systems.
- Migrated legacy on-prem batch pipelines to AWS (EC2, S3, Glue, Lambda), automating workflows to eliminate 60+ manual steps per week, improving data reliability and scalability.
- Developed API-based ingestion pipelines for multiple external vendors, integrating standardized error handling, retry logic, and monitoring, increasing data delivery success by 25% and reducing downstream reporting delays.
- Automated data validation and anomaly detection using Python, SQL, and PySpark, deploying 40+ scripts that identified over 90 potential production issues before deployment, safeguarding reporting quality.
- Containerized ETL workloads with Docker and integrated them into Jenkins CI/CD pipelines, accelerating deployment cycles and improving environment consistency.
- Collaborated with analytics and data science teams to prototype feature pipelines for regression and classification models, supporting exploratory ML initiatives on financial risk data.
- Implemented a GenAI proof-of-concept using OpenAI API and Snowflake, generating summarized insights for 15+ dashboards on quarterly financial trends, reducing manual reporting effort by 20 hours/month and enhancing internal dashboard adoption.
- Supported modernization of the data stack using Airflow, dbt, and Snowflake, improving pipeline observability, modularity, and governance to enable advanced analytics and AI/ML workflows.

Data Analyst | *Wipro, India* **Jan 2020 – Dec 2022**

- Automated data ingestion from 5+ external sources (APIs, flat files, Oracle) using Python scripts and schedulers, saving ~20 hours per week and improving data accuracy and timeliness for finance and operations reporting.
- Cleaned and migrated 1.6M+ records using Salesforce Data Loader and Python-based validation, removing over 12,000 duplicates and maintaining high-quality, reliable data for downstream reporting.
- Wrote and optimized 70+ SQL queries and stored procedures across Oracle and MySQL, performing transaction-level transformations and reducing query runtimes by up to 30%, ensuring faster access to actionable insights.
- Collaborated with data architects to design star schema models for enterprise reporting, reducing query runtimes by up to 25% and enabling faster, more accurate insights for business stakeholders.
- Created and maintained Tableau dashboards for 15+ stakeholders with hourly refresh cycles, enabling timely insights and informed decision-making across multiple business units.
- Leveraged Python for exploratory data analysis and anomaly detection, identifying trends and data inconsistencies that informed early-stage predictive modeling and operational improvements.
- Supported QA and release cycles in Agile teams, deploying biweekly reporting updates with minimal errors and ensuring smooth handoffs to analytics and business teams.

EDUCATION

Master of Science in Data Science
University of Massachusetts Dartmouth, MA, USA **Jan 2023 – Dec 2024**
Relevant Coursework & Graduate Projects: Data Warehousing, Cloud Data Engineering, Big Data Analytics

Bachelor of Technology in Information Technology
Vignana Bharathi Institute of Technology, Hyderabad, India **Aug 2017 – Aug 2021**

PROJECTS

NYC Taxi Data Analytics | Graduate Project (2023)

Technologies: *Azure Synapse, Python, scikit-learn, Power BI*

- Engineered a Synapse pipeline to process NYC taxi records, ensuring clean and reliable data for analysis.
- Built regression and clustering models to predict fares and identify high-demand zones, improving route efficiency.
- Developed interactive Power BI dashboards with borough and time filters to make insights more accessible.
- Automated preprocessing workflows in Python, streamlining the data preparation and model experimentation process.

Healthcare Claims Pipeline Automation with GenAI Insights | Capstone Project (2024)

Technologies: *Snowflake, Airflow, Jenkins, Tableau, Azure OpenAI, Vector Databases (Pinecone, FAISS), RAG Pipelines*

- Developed Airflow ETL pipelines to ingest, clean, and transform healthcare claims and encounter data for analysis in Snowflake.
- Automated pipeline scheduling and Tableau data refreshes using Jenkins CI/CD, improving reliability and repeatability.
- Created interactive dashboards to monitor claim status, denial patterns, and processing delays.
- Applied Azure OpenAI along with Vector Databases (Pinecone, FAISS) and RAG pipelines to generate summaries of anomalies and irregularities, demonstrating practical GenAI-assisted insights.

Customer Churn Prediction & Segmentation | AI/ML Project (2024)

Technologies: *Python, Pandas, scikit-learn, XGBoost, Snowflake, Power BI*

- Cleaned and engineered 500K+ customer records from Snowflake to build a unified dataset for churn modeling.
- Trained XGBoost models achieving 85% accuracy, enabling data-driven retention planning and customer outreach.
- Applied K-Means clustering to segment customers by behavior, helping marketing teams personalize campaigns.
- Built Power BI dashboards combining churn risk and segmentation insights, reducing attrition and improving targeting.
- Automated scoring and reporting pipelines using Python and Snowflake, creating a reusable predictive framework.

CERTIFICATIONS

- SnowPro Core Certification – Snowflake
- Oracle Cloud Infrastructure Data Platform Associate – Oracle
- Microsoft Azure Data Fundamentals (DP-900) – Microsoft
- Generative AI with Large Language Models – DeepLearning.AI (Coursera)
- dbt Fundamentals Badge – dbt Labs
- Snowflake Hands-on Labs (Project Badges) – Snowflake