

Bhushan Shelke

+1 317-507-1064 | bhushanshelke767@gmail.com | [linkedin.com/in/bhushan-shelke-iui](https://www.linkedin.com/in/bhushan-shelke-iui) | Little Rock, Arkansas

SUMMARY

Data Engineer with 2+ years of experience building scalable ETL pipelines, data models, and machine learning-driven data products on large-scale datasets. Skilled in Python (PySpark), SQL, and modern data engineering practices, including data orchestration, pipeline optimization, and cloud-based data processing. Experienced in integrating analytics, dashboards, and ML workflows to improve data accessibility, quality, and performance across distributed systems.

EDUCATION

Indiana University, Indianapolis	08/2023 – 05/2025
Master of Science in Applied Data Science (Statistics)	GPA: 3.8
University of Mumbai, Mumbai	08/2019 – 05/2023
Bachelor of Engineering in Electronics and Telecom	GPA: 3.7

WORK/INTERNSHIP EXPERIENCE

COSMOS Lab – University of Arkansas, USA Data Engineer I	08/2025 – Present
- Built a full-stack narrative intelligence platform (BlogTracker 2.0) analyzing 120K+ posts across 98 blogs, enabling real-time exploration of sentiment, topics, and influence - Analyzed millions of large-scale records using advanced SQL on PostgreSQL, identifying performance trends and optimization opportunities that informed data-driven recommendations - Designed data processing pipelines and data models for large-scale text datasets, enabling efficient storage, retrieval, and downstream analytics across distributed systems - Developed backend services (FastAPI) to expose processed data and model outputs for downstream applications and dashboards - Designed pipeline workflows with task sequencing and dependency management concepts similar to orchestration tools (e.g., Dagster/Airflow), improving pipeline reliability and maintainability - Developed scalable ETL pipelines using Python and SQL for processing large-scale datasets, incorporating modular design patterns similar to modern data stack practices (e.g., orchestration, transformation layers) - Conducted exploratory data analysis (EDA) and hypothesis-driven investigations using Python and SQL to identify trends, validate data quality, and surface actionable insights - Improved data accessibility and quality by implementing validation checks, schema consistency enforcement, and reliable data delivery mechanisms for downstream users - Integrated machine learning and AI pipelines (LLMs, clustering, embeddings) into data workflows, enabling advanced analytics and intelligent data products - Processed and analyzed large-scale datasets (120K+ records) using optimized batch processing, indexing, and vectorized operations for performance scalability	
Comet Lab, USA Data Scientist	11/2023 – 05/2025
- Developed and maintained ETL pipelines using Python and SQL to ingest, transform, and integrate data from APIs, analytics platforms, and structured sources - Designed structured data models to support analytics and dashboarding use cases - Designed and ran A/B tests and hypothesis-driven analyses to evaluate user engagement strategies - Developed statistical models to analyze user behavior, improving campaign tracking accuracy by 40% - Built dashboards (Tableau, Looker Studio) powered by structured data models, enabling self-service analytics on integrated datasets - Automated reporting workflows, reducing manual effort by 60% and enabling faster experimentation cycles - Partnered with product and UX teams to translate analytical findings into actionable changes	
Favorite Healthcare Staffing, USA Data Scientist	05/2024 – 01/2025
- Built ML models (XGBoost, LightGBM, logistic regression) to rank and recommend healthcare workers, improving matching accuracy - Developed feature engineering pipelines to transform raw data into model-ready datasets - Ensured data integrity through validation checks and preprocessing pipelines for reliable model training and inference - Evaluated bias in ranking models (e.g., leave-based bias) and applied statistical techniques to mitigate fairness issues - Conducted causal analysis and hypothesis testing to validate workforce allocation strategies - Delivered insights to stakeholders through SQL analysis plus reporting, influencing operational decisions	

SKILLS & TOOLS

Programming: Python (Pandas, NumPy, PySpark), SQL

Machine Learning: Regression, Classification, Clustering, LDA, Gradient Boosting

Data Engineering: ETL Pipelines, Data Modeling, Data Processing, Data Validation, Pipeline Optimization

Data Systems: PostgreSQL, BigQuery, ETL Pipelines, Data Modeling

Visualization: Tableau, Power BI, Looker Studio

LLM / AI: LLM evaluation, embeddings, similarity scoring, prompt pipelines

Cloud & Tools: AWS (S3, EC2 – basic), Azure (Data Factory, Power BI – exposure), Docker (basic), Git