Vamsidhar Boddu

Data Engineer | ML Engineer | Data Scientist

Richmond, VA | +1 (571)-405-0806 | vamsidharboddu16@gmail.com | LinkedIn

SUMMARY

Data Engineer with **4+ years of experience** delivering enterprise-scale data platforms across education, energy, finance, and academic research, building high-performance ETL/ELT pipelines on AWS, Azure, and Databricks processing 10M–15M+ records daily. Proven track record improving analytical reporting speed by 40%+, increasing ML model readiness by 30%, and reducing data quality issues by 35–50% through PySpark, Delta Lake, Airflow, Snowflake, and Great Expectations. Successfully supported real-time fraud analytics, student retention intelligence, IoT refinery monitoring, and lending risk engines consistently driving measurable business outcomes with precision, governance, and automation excellence.

SKILLS

Programming & Data Processing: Python (Pandas, NumPy, Scikit-learn, PyTorch, TensorFlow), SQL, PySpark, Spark Structured Streaming, Shell Scripting, Databricks, Scala (basic), RESTful API Integration
Data Engineering & ETL: End-to-end ETL/ELT pipeline design, data modeling (Star/Snowflake), transformation logic, orchestration using Apache Airflow, AWS Glue, Azure Data Factory (ADF), Spark SQL optimization, CDC processing, data partitioning & schema evolution
Databases & Storage Systems: MySQL, PostgreSQL, Snowflake, AWS Redshift, Azure SQL Database, BigQuery (exposure), MongoDB, Cassandra, Delta Lake, Parquet, ORC, HDFS
Machine Learning & Data Science: Regression, Classification, Clustering, XGBoost, Random Forest, LightGBM, Feature Engineering, Hyperparameter Tuning, Model Evaluation (ROC-AUC, Precision-Recall), Time Series Forecasting, A/B Testing, NLP (SpaCy, Hugging Face)
MLOps & Model Deployment: Model packaging & deployment via FastAPI/Flask, MLflow & DVC for model tracking & versioning, CI/CD pipelines for ML workflows (GitHub Actions, Jenkins), AWS SageMaker / Azure ML (exposure), real-time inference APIs, model drift detection
Data Quality & Governance: Great Expectations, Deequ, Data Validation Frameworks, Data Lineage & Provenance Tracking, Data Observability, SLA monitoring, Governance & Compliance Automation (HIPAA, GDPR)
Analytics & Visualization: Tableau, Power BI, Looker Studio, Plotly, Matplotlib, KPI Dashboards, Excel (Power Query, Pivot), SQL & Python-based BI Reporting, Forecasting Automation
Cloud & DevOps Enablement: AWS (S3, Glue, Lambda, Redshift, EMR), Azure (Data Factory, Data Lake, Synapse), GCP (BigQuery, Dataflow), Terraform (basic), Docker, Kubernetes (exposure), CI/CD pipelines, Git, Agile/Scrum collaboration
Version Control & Collaboration: Git, GitHub, GitLab, Bitbucket, JIRA, Confluence, Agile & CI/CD workflow integration

WORK EXPERIENCE

Data Engineer | EAB Global Inc., Richmond, VA **June 2025 – Present**

- Automated ingestion of 1M–1.5M student and alumni records daily from 10+ LMS connectors using AWS Glue, Lambda, and EMR (PySpark), achieving 99.9% uptime and maintaining a sub-5-minute SLA for data freshness.
- Designed multi-tier ingestion architecture integrating S3 (raw zone), RDS (metadata), and Redshift (DWH) on EKS/EC2 clusters, supporting 50K+ concurrent analytic queries for institutional dashboards and predictive student performance analytics.
- Developed Redshift star schemas with strategic partitioning and clustering, reducing query runtime from minutes to seconds and optimizing performance across 2,500+ universities using advanced indexing and workload management configurations.
- Implemented FERPA-compliant data governance through IAM role controls, schema validation rules, and dynamic PII masking, securing over 50TB of sensitive datasets across cloud environments with zero compliance violations.
- Delivered curated analytical datasets fueling Navigate360's ML pipelines, enabling real-time student retention predictions and advisor interventions that improved early risk detection accuracy by 24% across multi-campus institutions.

Cloud & Data Support Assistant | George Washington University, Washington, D.C. **August 2023 – May 2025**

- Developed automated ETL jobs using Python, Azure Data Factory, and SQL to migrate 5TB+ student services data securely across academic systems, improving processing efficiency by 38% and reducing manual data refresh dependencies.
- Built Power BI and Tableau dashboards tracking application trends, department performance, and research funding utilization, enabling data-driven decisions that increased student service request resolution speed by 29% across university units.
- Implemented automated data quality validation scripts using Great Expectations and Azure SQL procedures, reducing data discrepancies by 41% and ensuring compliance with institutional reporting and FERPA governance requirements.
- Supported faculty analytics initiatives by preparing ML-ready datasets in Databricks for enrollment behavior prediction models, accelerating experimentation workflows by 34% while ensuring secure access control across academic stakeholders.

- Coordinated with cross-functional academic teams to optimize data storage within Azure Data Lake and Snowflake, reducing query execution time by 45% through partition strategy refinement and schema normalization best practices.

Programmer Analyst / Data Engineer | Cognizant Technology (Client: PETRONAS), India **February 2022 – July 2023**

- Engineered PySpark and Azure Data Factory pipelines to consolidate 10M+ daily IoT sensor records from refinery operations, improving near real-time monitoring accuracy by 37% for downstream equipment failure prediction systems.
- Optimized Delta Lake and Azure Synapse data models through partition pruning and Z-ordering, reducing query execution time by 45% and accelerating business intelligence availability for drilling performance analytics teams.
- Implemented automated data quality rules and anomaly alerts using Great Expectations and Databricks notebooks, cutting data reconciliation effort by 50% and ensuring accurate reporting for production throughput and asset health KPIs.
- Collaborated with data science team to deliver feature-engineered datasets for predictive maintenance ML models, improving model training efficiency by 33% and increasing anomaly detection precision for high-value refinery assets.
- Deployed CI/CD workflows using Git, Azure DevOps pipelines, and unit-tested Spark transformations, ensuring version-controlled deployments and reducing data pipeline rollback incidents by 28% across production environments.

Data Analyst | DMI Finance, India **January 2021 – January 2022**

- Developed SQL and Python-based credit risk analysis workflows processing 5M+ loan application records monthly, improving default flagging accuracy by 31% and accelerating underwriting decision support for consumer lending products.
- Built actionable Power BI dashboards tracking EMI performance, customer delinquency trends, and portfolio segmentation, enabling business teams to increase early-stage recovery targeting efficiency by 28% across high-risk cohorts.
- Automated ETL processes using Python, SSIS, and PostgreSQL to centralize fragmented financial source feeds, reducing manual reporting preparation effort by 45% and improving data readiness for regulatory audit submissions.
- Conducted A/B analytics on borrower behavior trends using statistical modeling and segmentation logic, improving credit scoring rule optimization by 22% and lowering false declines within low-risk income groups.
- Partnered with risk and legal stakeholders to enhance data governance workflows by defining validation checkpoints, decreasing mislabeled loan records by 35% and ensuring compliance with NBFC reporting frameworks.

EDUCATION

Master of Science in Data Science

George Washington University, Washington, D.C., USA **Aug 2023 – May 2025**

Certifications

- **AWS Certified Cloud Practitioner**
- **AWS Certified Data Engineer – Associate (DEA-C01)**
- **Microsoft Certified: Azure Fundamentals**
- **IBM Data Science Foundations – Level 1**

PROJECTS

Predicting EV Adoption & Charging Infrastructure | Washington State

- Forecasted 1.2M+ EV adoption records using Facebook Prophet with 92% accuracy to project statewide trends through 2028.
- Mapped demand-supply gaps using geospatial SQL & Dash dashboards, identifying 2,500+ priority charger locations to support WA's 2035 zero-emission mandate.

TransGuard360 – Real-Time Fraud Analytics Pipeline

- Engineered Kafka + Spark Streaming pipeline handling 10M+ financial transactions daily with <5s latency and fault-tolerant scaling.
- Automated ETL workflows via Apache Airflow and optimized fraud detection using SQL + Amazon QuickSight, boosting accuracy by 35% and reducing false positives by 40%.

Mental Health Risk Prediction using NLP (TensorFlow)

- Developed TensorFlow-based NLP classifier analyzing 50K+ patient statements with 92%+ accuracy to predict clinical mental health risk categories.
- Cut diagnostic analysis time by 50% and reduced model training runtime by 35%, validated across 100+ real-world anonymized patient case records.

AutoShip – CI/CD Automation with Jenkins & AWS EKS

- Built end-to-end CI/CD pipeline using Jenkins, Docker, Ansible, and Amazon EKS to automate zero-downtime Kubernetes deployments.
- Eliminated 80% manual deployment effort while supporting 100K+ daily API requests with 99.99% uptime through blue-green rollout strategy.