

VINEETH KUMAR DEEPALA

Phone: +1 (919) 505-6970 • Location: Cary, NC • Email: vineethkumard99@gmail.com • [LinkedIn](#) • [GitHub](#)

SUMMARY

Data Engineer with 3+ years specializing in Databricks/Spark, Snowflake, SQL, and AWS. Improved downstream data availability, cut job runtimes, and reduced compute costs. Builds ML-ready feature stores, automated anomaly detection, and produces clear, actionable data visualizations (Tableau, Power BI) that translate complex signals into operational insights for stakeholders. Doctoral candidate in Computer Science (distributed systems and machine learning); on F-1 visa with approved cap-subject H-1B petition (Form I-797B), cap-exempt for employer transfer.

TECHNICAL SKILLS

- **Programming:** Python, C++, Java, SQL, Shell
- **Distributed Systems:** Apache Spark, Kafka, Airflow, Hadoop, AWS (S3, EMR), GCP (BigQuery, Dataflow), Informatica, Splunk, Snowflake, DuckDB, dbt, FastAPI, CI/CD
- **Machine Learning:** TensorFlow, PyTorch, scikit-learn, CNNs, RNNs, LSTMs, ML pipelines
- **Statistical & Anomaly Detection:** Statistical analysis, time-series modeling, anomaly detection, feature drift detection
- **Visualization:** Power BI, Tableau, Streamlit
- **Tools:** Git, Docker, Linux/Unix

EDUCATION

University of the Potomac - Washington, DC

Doctoral in Computer Science (Expected 2029)

University of New Haven - West Haven, CT

M.S. in Data Science (Dec 2022) | GPA: 3.34

RELATED EXPERIENCE

Sanizer IT Services - Data Engineer (Contract)

Texas, USA | Sep 2025 - Dec 2025

- Designed and optimized low-latency Databricks/Spark streaming pipelines to ingest high-throughput event streams for real-time ML anomaly detection, improving ingestion stability and reducing processing.
- Applied memory and execution-plan optimizations (partitioning, serialization, caching) to eliminate bottlenecks and reduce task failures
- Built an ML-ready feature store on AWS S3 and Snowflake with versioning and automated feature pipelines for reproducible experiments and faster model iteration.
- Integrated statistical drift and anomaly-detection checks into Databricks workflows, enabling automated alerts and improving monitoring precision.
- Executed metadata and lineage tracking using AWS Glue and created Tableau dashboards to surface feature drift and model health to stakeholders.

Nationwide Insurance - Data Engineer (Contract)

Ohio, USA | June 2023 - July 2025

- Optimized batch and streaming data pipelines in Databricks using PySpark and SQL, ingesting claims and policy data from Guidewire ClaimCenter and PolicyCenter via DataHub, improving downstream data availability by 30%.
- Migrated six legacy Informatica and Perl-based batch workflows to Databricks Lakehouse, reducing average job runtime by ~25% through Spark-based parallel processing.
- Designed an automated data quality framework (schema validation, null/duplicate detection, reconciliation checks) across 12 tables, cutting business reported data incidents.
- Implemented pipeline monitoring and alerting using Splunk and New Relic, cutting mean time to detect pipeline failure by 30%.
- Developed Power BI reports to track daily data quality metrics and schema validation failures, enabling business stakeholders to monitor pipeline reliability and prioritize fixes.

Anthem - Data Engineer (Contract)

Texas, USA | Jan 2023 – May 2023

- Engineered high-throughput ETL pipelines processing 200M+ healthcare records monthly using Spark and Python with optimized memory and execution plans.
- Designed HIPAA-compliant data lakehouse architecture on AWS with encryption, access control, and audit logging for reliable monitoring.
- Implemented data masking for PII (patient names, SSNs, medical record numbers) in Python extraction scripts, applied at the ingestion layer to prevent exposure in logs, temporary tables, and query results.
- Built multithreaded validation framework detecting 1,200+ anomalies per cycle using statistical profiling and rule-based checks.
- Created HIPAA compliance documentation, including data flow diagrams, risk assessments, and technical specifications for external audit teams and compliance review.
- Tuned Snowflake performance (clustering, materialized views, caching), reducing compute cost by 20% while improving query latency.

CloudVik LLC – Data Engineer Intern (Internship)

Texas, USA | Summer 2022

- Gained hands-on experience with Hadoop ecosystem components and distributed data processing.
- Migrated datasets to AWS and GCP environments for scalable analytics.
- Collaborated with team members using SQL and Python to prepare and filter data for cloud storage and processing.
- Designed basic Power BI dashboards to validate data filters and summarize dataset metrics before migrating them to AWS and GCP environments

RESEARCH & ACADEMIC PROJECTS

FoodFlow: Intelligent Food Waste Lakehouse & SLA Decision Engine

- Architected a 3-tier Medallion lakehouse using DuckDB and dbt to process IoT disposal streams and retail transactions, aggregating daily waste summaries against EPA benchmarks.
- Built a sub-15ms FastAPI microservice providing collaborative-filtering recommendations constrained by USDA FoodKeeper shelf-life rules (≤ 3 days suppression) and RAG-assisted anomaly diagnostics.
- Created an automated CI/CD pipeline via GitHub Actions and Docker, ensuring schema consistency, cold-start safety, and 100% test coverage on Ubuntu runners.
- Developed an interactive Streamlit operations dashboard connecting directly to gold fact marts and microservice endpoints for real-time monitoring

Automatic Image Captioning Using Deep Learning and NLP

- Build CNN-LSTM pipeline (VGG16 + LSTM) for image captioning on Flickr8k dataset, improving BLEU score by 15%.
- Designed preprocessing pipeline for tokenization, embeddings and vocabulary constructions.
- Applied transfer learning, batch normalization and dropout to improve generalization.

RetinaFace: Single-stage Dense Face Localization

- Implemented face detection using ResNet-152 and MobileNet with multi-task learning for bounding boxes and landmarks.
- Built a custom annotation and augmentation pipeline; optimized training with LR scheduling and gradient clipping.
- Reduced localization loss by 40% across 10 epochs.

PUBLICATIONS

Deepala, V.K., Buttul, S.F., & Pally, C. (2025). *Generation of Automatic Image Captioning Using Natural Language Processing*.