Kranthi Chaithanya Thota

Chapel Hill, NC | +1 (315) 603-7581 | kranthichaithanya3000@gmail.com | LinkedIn | kc-3000.com

Summary

Data and GIS Engineer with 4 years of industry experience building cloud-native data platforms, geospatial pipelines, and AI/ML systems. Deep expertise in GCP, Python, SQL, and PostgreSQL. Proven track record delivering scalable ETL/ELT pipelines, real-time streaming architectures, and spatial analytics across government, infrastructure, and enterprise domains. Experienced with LLM/RAG systems, computer vision, and NLP.

Certifications

- Google Cloud Professional Data Engineer
- Google Cloud Associate Cloud Engineer
- HashiCorp Certified Terraform Associate

Professional Experience

Data Engineer | University of North Carolina at Chapel Hill, Chapel Hill, NC Sep 2025 – Present

- Engineered a **PySpark** feature pipeline on **AWS EMR** over **50M+** rows using broadcast joins and dynamic partition pruning to cut job runtime by **55%**; persisted outputs to **Delta Lake** with Z-ordering, reducing downstream ML feature query scan cost by **70%**.
- Architected a containerized microservice for unstructured data ingestion (OCR), designing a template-agnostic schema generalizing across **300+ document formats**; implemented fault-tolerant retry logic achieving **95% reliability**.
- Engineered a **Star Schema** data warehouse consolidating **5+ heterogeneous federal datasets**; reduced average query runtime by **60%** via materialized views, enabling low-latency analytics for downstream reporting.
- Implemented an automated data observability framework using statistical drift detection, reducing downstream data cleaning latency by **40%** by catching schema violations at ingestion time.
- Refactored legacy monolithic scripts into modular Python applications deployed on **Kubernetes**; established CI/CD pipelines enabling zero-downtime deployments.

Data Engineer Intern | Clarkson CEM Consulting Group, Potsdam, NY May 2025 – Aug 2025

- Built and fine-tuned a **YOLOv8 + OpenCV** computer vision pipeline to analyze UAV drone footage, achieving **95% accuracy** in real-time parking occupancy detection; results informed campus safety resource allocation with a projected **20% efficiency improvement**.
- Designed and implemented scalable ETL pipelines processing **6+ years** of classroom and facilities scheduling data; identified underutilized resources and scheduling inefficiencies, saving an estimated **200+ hours** of manual effort annually.
- Conducted roadkill hotspot analysis using **ArcGIS Pro** and spatial statistical methods, identifying critical corridors and delivering actionable geospatial findings to stakeholders.

Graduate Research & Teaching Assistant | Clarkson University, Potsdam, NY Feb 2024 – Aug 2025

- Designed and delivered graduate-level coursework on Data Warehousing and Relational Database Systems covering schema design, normalization, indexing, and query optimization using **SQL, PostgreSQL, and Snowflake**; mentored **50+ students**.
- Engineered an automated data pipeline using **BeautifulSoup** to scrape 10,000+ **NYSERDA** grant records and a fine-tuned Hugging Face Transformer for Named Entity Recognition, achieving 90% F1-score across organization, funding, and project entities.
- Automated migration of **20+ years** of legacy IPEDS data into a centralized data warehouse; developed **30+ KPI dashboards** benchmarking peer institutions, reducing manual data retrieval time by **60%**.
- Engineered a **Power BI** analytics platform with advanced forecasting models, enabling self-service insights across **10+ departments** and **500+ employees**, saving **150+ analyst hours** annually.

Associate Data Engineer | Egen (SpringML) , Hyderabad, India Jan 2022 – Dec 2023

- Engineered an end-to-end geospatial data pipeline for **NY Power Authority (NYPA)** EV charger site planning, acquiring and processing **10+ public GIS datasets** using **QGIS, Python, BigQuery, and Dataflow**; built a scoring algorithm evaluating location suitability and developed backend APIs powering a map-based decision tool used by government stakeholders.
- Architected a medallion data warehouse in **Snowflake** using **dbt** (40+ models, schema tests, source freshness checks gated in GitHub Actions CI); implemented Streams and Tasks for **CDC**, reducing ELT latency from hourly batch to **sub-5-minute** incremental loads.
- Built a scalable **Document AI** pipeline ingesting **2,000+ patent documents** with ongoing daily ingestion of 50+ new documents, generating **25+ KPIs** deployed via Cloud Run; reduced manual patent review time by **80%**.
- Optimized multi-stage **Apache Airflow** DAGs via dynamic task generation, reducing end-to-end pipeline runtime by **33% (6h to 4h)** and improving production data reliability.
- Automated provisioning of **50+ GCP resources** (BigQuery, GCS, Compute Engine, Cloud Run) using **Terraform** with CI/CD via Cloud Build, reducing environment setup time by **70%** and enforcing security and compliance controls.
- Built backend APIs and internal data tools using **Flask** and Python, including SSO-authenticated web applications and cross-team analytics platforms integrating Google Workspace data into BigQuery.

Technical Skills

Languages: Python, SQL, R, JavaScript (Angular, D3.js), C++, Go
Cloud & Infrastructure: GCP (BigQuery, Dataflow, Cloud Run, Pub/Sub, Vertex AI, Apigee, Cloud Scheduler), AWS (EMR, Redshift, Glue, Lambda, S3, SageMaker), Terraform (IaC), Docker, Kubernetes
Data Engineering: Apache Spark (PySpark), Apache Kafka, Apache Airflow, dbt, Snowflake, Delta Lake, ETL/ELT Design, Star Schema / Medallion Architecture, CDC, Data Modeling, REST API Integration
Databases: PostgreSQL (triggers, PL/pgSQL, encryption, raw SQL ETL), BigQuery, Snowflake, Redshift, MongoDB, MySQL, TimescaleDB, SQL Server
GIS & Geospatial: ArcGIS Pro, ArcPy, QGIS, PostGIS, Spatial SQL, Network Analysis, Rasterization, Georeferencing, Layer Classification, Deep Learning for GIS
ML & AI: PyTorch, TensorFlow, Transformers (BERT, LLMs, RAG), NLP, Computer Vision (YOLOv8, OpenCV, EfficientNet), Reinforcement Learning, Hugging Face, Ollama, OpenAI API
DevOps & Visualization: CI/CD, GitHub Actions, Cloud Build, Jenkins, Git, Tableau, Power BI, Looker, Grafana, Matplotlib, Seaborn, Plotly

Projects

Real-Time Reddit Stock Sentiment Tracker *Python, TimescaleDB, Reddit API*

- Engineered a low-latency Python streaming pipeline processing **100+ comments/sec** using tumbling window logic to surface breakout stock tickers in under **5 seconds**; reduced trend detection latency by **95%** vs. batch methods via TimescaleDB time-series storage.

HAVK Mladost Sports Club Data Infrastructure *PostgreSQL, Python, Looker Studio*

- Designed a normalized **PostgreSQL** schema with **15+ interrelated tables** and automated ingestion via Python ETL pipeline, replacing fragmented Excel workflows; reduced manual data handling by **95%** and improved membership renewal rates by **30%** through Looker Studio KPI dashboards.

Serverless Data Job Deployment Framework *Terraform, GCP, GitHub Actions*

- Architected a reusable IaC framework cutting new data job deployment time by **80%** (3+ hours to under 10 minutes) across **3 teams**, with standardized IAM roles, secret management, and compliance enforcement built directly into Terraform modules.

Education

M.S. Applied Data Science | Clarkson University, Potsdam, NY Jan 2024 – Aug 2025
GPA: 4.0/4.0 | Data Warehousing, Big Data Architecture, Cloud Computing, Data Mining, GIS & Spatial Analysis