Kranthi C. Thota

9196369363 kranthichaithanyathota@gmail.com linkedin.com/in/kranthi-c-thota kc-3000.com

Education

Clarkson University Potsdam, USA
Master of Science – Applied Data Science; CGPA: 4.0/4.0 *Jan 2024 – Aug 2025*

Professional Experience

Environmental Finance Center (EFC) at University of North Carolina Chapel Hill, NC
Data Engineer *Sep 2025 – Present*

- Engineered a three-tier PDF extraction pipeline (Camelot lattice to stream to Qwen2.5-7B LLM fallback with vision-layer cross-validation) processing 300+ heterogeneous CIP report formats; cut analyst extraction time by 85% and eliminated 700+ manual data-entry hours annually across 4 research teams.
- Architected a federated data lakehouse on AWS (S3 + Apache Iceberg) consolidating 8 federal water-sector datasets (400 GB) with systemized schema evolution, partition pruning, and column-level access controls; reduced p99 analytical query latency by 72% across 3 concurrent workloads.
- Designed a Star Schema warehouse on Amazon Redshift with pre-computed materialized views, cutting average analyst query time from 50 s to 9 s (82%) for dashboards consumed daily by 20+ Division of Water Infrastructure stakeholders.
- Implemented GitOps CI/CD for 40+ pipelines, dbt models, and SQL objects via GitHub Actions; shortened deployments from 3 days to 4 hours and decreased production incidents by 65% through automated lint, unit-test, and rollback gates.

Clarkson CEM Group Potsdam, NY
Data Engineer *May 2025 – Aug 2025*

- Built a real-time data quality monitoring framework with Apache Flink (Java) tracking schema drift, null-rate spikes, and throughput anomalies across 1,000+ Kafka topics; integrated PagerDuty alerting and lowered mean time to detect (MTTD) producer errors from 50 min to under 6 min, preventing 4 end-customer SLA breaches.
- Engineered a Kafka consumer lag forecasting microservice using Python, Prophet, and TimescaleDB that predicted consumer group saturation 25 minutes ahead with 84% accuracy across 350+ consumer groups, enabling proactive cluster auto-scaling before user-visible degradation.
- Developed a reusable Flink job deployment template on GKE with configurable parallelism, tunable checkpointing, and exactly-once state semantics; adopted by 3 internal teams and cut new Flink job onboarding from 2 days to under 3 hours.
- Optimized incremental checkpointing for a stateful Flink job handling 80K events/sec, reducing checkpoint overhead by 35% and improving sustained throughput by 18% with no compromise to fault-tolerance guarantees.

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

- Enabled statewide EV charger site selection for New York Power Authority (NYPA) by engineering a geospatial pipeline processing 15+ GIS datasets (50M+ spatial records) via QGIS, BigQuery, and GCP Dataflow; delivered a location-scoring API and interactive planning tool used by NYPA planners to identify 3,400 optimal charging sites across New York State.
- Accelerated client analytics from weekly batch to near-real-time by architecting a Snowflake warehouse with 50+ dbt models (staging, intermediate, mart) powered by CDC via Debezium and Kafka; reduced data freshness lag from 6 hours to under 3 minutes and cut stale-data incidents by 90% for a 120-user BI platform.
- Eliminated a manual patent extraction workflow on a GCP DocAI pipeline (Cloud Run + BigQuery) processing 500K+ documents annually at 94% accuracy; fed 6 live client dashboards and removed 80+ hours/month of analyst work.
- Reduced Airflow DAG runtime by 40% (6 hrs to 3.6 hrs) and cut P1 pipeline failures by 65% by refactoring 20+ legacy DAGs to dynamic task mapping, adding idempotent retry logic, dead-letter-queue alerting, and upstream dependency gating.
- Automated cloud provisioning for 60+ resources across 5 projects using Terraform, cutting setup time from 4+ hours to 28 minutes (88% reduction) with audit-ready IAM roles, VPC Service Controls, and organization policy guardrails; Designed modular, reusable IaC templates with clearly defined, minimal modification points for additional projects and environments.

Certifications

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

Skills

Programming & Scripting: Python, SQL, PySpark, JavaScript, PowerShell, Bash
Cloud & Infrastructure: GCP (BigQuery, Dataflow, Cloud Run, Pub/Sub, Vertex AI), AWS (EMR, Redshift, Glue, Lambda, S3), Terraform, Docker, Kubernetes
Data Engineering: Data Warehousing, Kafka, Delta Lake, Pub/Sub, ETL/ELT Development, Data Modeling, API Integration.
GenAI & LLMs: PyTorch, HF Transformers, LangChain, RAG pipelines,LLM Fine-tuning
DevOps & Version Control: Git, GitHub, Azure DevOps, CI/CD Pipelines
GIS & Databases: ArcGIS Pro, QGIS, PostGIS, PostgreSQL, MongoDB, SqlServer, MySQL, TimescaleDB