SAI KIRAN ETIKAPALLY

Dallas, TX • 331-226-3849 • kirane8989@gmail.com • LinkedIn • GitHub

PROFESSIONAL SUMMARY

Senior GCP Data Engineer with 9+ years designing enterprise-scale data platforms for Fortune 100 and healthcare clients. Deep expertise in BigQuery, Cloud Composer (Airflow), Dataproc (PySpark/Scala), and Python-based ETL/ELT engineering. Led multiple end-to-end Teradata-to-GCP migrations, automated UAT reconciliation workflows, and built cost-optimized, HIPAA-compliant data pipelines processing multi-TB workloads daily. Strong software engineering practices: CI/CD, automated data quality, IaC (Terraform), and version-controlled pipeline development.

CORE TECHNICAL SKILLS

Cloud Platforms: GCP (BigQuery, Dataproc, Cloud Composer/Airflow, GCS, Pub/Sub, Dataflow, Vertex AI), Azure (ADF, Databricks, ADLS, Azure SQL DWH)

Data Engineering: Apache Spark (PySpark, Scala), dbt, Delta Lake, Apache Kafka, Apache Hive, Sqoop, Informatica PowerCenter

Languages: Python, SQL, Scala, Shell Scripting (Unix/Linux), Spark SQL

Databases & Warehouses: BigQuery, Snowflake, Teradata, PostgreSQL, SQL Server, Oracle, MongoDB

Data Quality & Observability: Automated reconciliation frameworks, Great Expectations, Cloud Monitoring, Dataplex

DevOps & IaC: Git/GitHub, GitHub Actions, Cloud Build, Terraform (GCP), CI/CD pipelines, Docker

Visualization: Tableau, Looker, Power BI, Qlik

Certifications: Google Certified Professional Data Engineer | Databricks Lakehouse Fundamentals

PROFESSIONAL EXPERIENCE

GCP Data Engineer | Verizon | Irving, TX | Mar 2025 – Present

- Led end-to-end Teradata → BigQuery migration for enterprise Business Markets reporting, migrating 20 TB of historical data across 250+ production tables with zero data loss and minimal downtime.
- Designed and deployed 65+ Cloud Composer (Airflow) DAGs orchestrating multi-stage ETL pipelines processing 4 TB/day of enterprise customer, campaign, and funnel data into BigQuery.
- Built a Python-based automated reconciliation framework that reduced manual UAT validation time by ~80%, cutting validation cycles from 3 days to 4 hours.
- Optimized BigQuery tables via partition pruning, clustering, and materialized views, reducing average query latency by 45% and monthly compute spend by ~$18,000.
- Re-architected legacy spool-heavy Teradata models into a normalized BigQuery layer, eliminating 80+ ad-hoc spool tables and stabilizing 150+ downstream Qlik/Tableau reports.
- Implemented automated data quality checks across 12 pipeline stages using Python, reducing downstream reporting discrepancies by 85%.
- Configured Zenalytics YAML-based semantic layer definitions for 22 marketing analytics velocity queries, reducing new query deployment time from 3 days to 4 hours.
- Authored CI/CD-enabled DAG deployment workflows using GitHub Actions and Cloud Build, enabling automated testing and zero-downtime pipeline updates.
- Automated migration of 150+ legacy Teradata scripts and stored procedures to BigQuery SQL using LLM-assisted translation — initially piloted with open-source LLMs for SQL dialect conversion, then migrated to Claude AI (Anthropic) for higher accuracy and contextual optimization; reduced manual rewrite effort by ~75%, cutting per-sprint migration cycle from 6 weeks to under 2 weeks.

GCP Data Engineer | Tenet Healthcare | Dallas, TX | Mar 2023 – Feb 2025

- Architected HIPAA-compliant GCP data infrastructure processing 1.5 TB/day of patient and operational data across 55 hospital systems, maintaining 99.9%+ pipeline uptime.
- Led Teradata-to-BigQuery migration for clinical and operational datasets; rebuilt 120+ Informatica ETL mappings as Cloud Composer DAGs and BigQuery SQL, cutting pipeline runtime by 35%.
- Optimized PySpark jobs on Dataproc using partitioning, broadcast joins, and caching, reducing large patient dataset processing time by 40% and cutting Dataproc cluster costs.
- Built a scalable GCP data lake (GCS + BigQuery + Snowflake) enabling 10 analytics teams to self-serve query clinical data, reducing time-to-insight from 5 days to 4 hours.
- Automated data reconciliation and PHI-compliant validation using Python scripts, reducing manual QA by 20 hours/week and catching data anomalies before downstream impact.
- Enforced PHI/PII data governance via GCP IAM RBAC and DLP API integration across all production pipelines in compliance with HIPAA and GDPR requirements.

GCP Data Engineer | DriveWealth | New York, NY | Jan 2021 – Feb 2023

- Built end-to-end financial data ingestion pipelines in Cloud Composer processing 15 million daily transactions from 6 upstream sources into BigQuery and GCS with 99%+ SLA compliance.
- Developed PySpark transformation jobs for multi-format financial data, enabling real-time customer usage aggregations at 3+ billion-row scale.
- Automated data cleansing and validation workflows using Python, decreasing data preparation time by 50% and improving pipeline reliability to 99.5% uptime.
- Designed Snowflake data warehousing layer integrated with GCP BigQuery, supporting financial reporting and BI for 12 downstream consumers.

Azure Data Engineer | TJX Companies | Framingham, MA | Jul 2019 – Dec 2020

- Designed ADF pipelines migrating 8 TB of retail transaction data from SQL Server to Azure Data Lake and Azure SQL Data Warehouse.
- Built PySpark ETL jobs on Azure Databricks with Delta Lake for real-time inventory and sales analytics; implemented Airflow DAGs for automated Snowflake data loads.
- Developed RESTful Python (Flask/Django) APIs integrating diverse upstream data sources into the enterprise analytics platform.

Hadoop Developer | Evoke Technologies | India | Oct 2016 – Oct 2018

- Built MapReduce, Hive, and PySpark data processing pipelines on Hadoop/HDFS; designed HBase schemas and automated data ingestion from MySQL via Sqoop.

EDUCATION

Master of Science, Information Science | Trine University, Angola, IN | 2022