AMRITA NEOGI

(520) 427 - 1767 | Austin, TX | neogiamrita111@gmail.com | http://linkedin.com/in/amritaneogi

SUMMARY

Data Professional with 6.5+ years building ETL pipelines in healthcare (Epic/EHR, OMOP) and finance, specializing in BigQuery design, data governance, and ML-ready pipelines on GCP. TCS *Contextual Master* awardee for integrating 2.5M customer records with Protegrity tokenization and achieving a 30% processing gain.

WORK EXPERIENCE

Data Manager, ARID Lab, University of Arizona, Tucson, AZ — **Feb 2024 – Oct 2025**

- Engineered end-to-end ETL pipelines in SQL and Python on BigQuery and PostgreSQL to ingest and standardize Epic EHR data from 10+ facilities into OMOP CDM, using partitioning, clustering, and deduplication to reduce data prep time by 60%.
- Designed dimensional models and ML-ready datasets for survival and utilization forecasting, delivering 15+ clinical KPIs and Tableau dashboards that identified high-risk cohorts with 83% prediction accuracy.
- Automated data quality and governance workflows (schema validation, completeness checks, outlier detection, lineage) and built standardized data dictionaries to ensure HIPAA-compliant, audit-ready datasets for compliance reporting.
- Built Python NLP pipelines to extract diagnosis, medication, and lab fields from unstructured clinical notes, improving accuracy by 25%, reducing manual review by 70%, and mentoring three graduate students in data engineering best practices.

Graduate Research Assistant, Dept. of Pediatrics, University of Arizona, Tucson, AZ — **Nov 2022 – Dec 2023**

- Optimized PostgreSQL databases for multi-site Medicaid and public-health analytics, applying indexing, query tuning, and batching strategies that reduced runtimes by 80% and improved data completeness to 92.5% across regional datasets.
- Built automated validation and ingestion workflows connecting REDCap and AWS Athena, adding rule-based QA checks and role-based access controls to support secure, high-quality clinical research data collection.
- Developed AWS-based ETL pipelines to load raw state vital records into S3, transform them with Athena SQL and ICD-code mappings, and produce partitioned, OMOP-aligned analytical tables for mortality trend and anomaly analysis.

Software Development Engineer, Tata Consultancy Services (TCS), India — **Mar 2018 – Jul 2022**

- Led a 12-member team building Informatica and pySpark ETL pipelines processing 10M+ daily banking transactions from mainframe to Teradata, implementing Protegrity PII tokenization, unit testing, and source-to-target reconciliation.
- Designed and optimized large-scale data pipelines supporting fraud and risk analytics, reducing false positives by 20% through statistical checks, anomaly investigation, and trend analysis.
- Automated Git-based CI/CD for ETL releases, integrating pySpark batch jobs with data quality gates, rollback controls, and standardized deployments that improved release cadence by 25%.
- Ensured regulatory compliance and data integrity across 3M+ daily transactions through controlled ingestion patterns, audit logging, and validation frameworks supporting enterprise risk reporting.

PROJECTS

Healthcare Outcomes & Risk Analytics — *R/SAS • SQL • Tableau • Statistical Modeling*

- Analyzed a 114K-infant dataset using R/SAS & SQL pipelines to evaluate survival differences by payer type.
- Identified ~2.65x higher mortality risk for uninsured infants and summarized findings in Tableau for clinical and policy teams.

Healthcare Utilization & Guideline Adherence — *SQL • Python • Survival & Clustering • Tableau*

- Analyzed 50K+ claims using survival/clustering models and quality checks to assess utilization and guideline adherence.
- Delivered Tableau summaries showing three utilization profiles with AUROC above 0.85 for planning and policy use.

Regulatory Risk Analytics — *Python • Airflow • Data Quality • Forecast Modeling*

- Automated regulatory risk reporting using Python & Airflow with forecasting models and quality checks, improving accuracy
- Reduced manual effort by 75%, kept forecast error under 15% and enabled proactive, data-driven compliance planning.

EDUCATION

MS in Data Science, University of Arizona, Tucson, AZ — **Aug 2022 – Dec 2023**
GPA: 4.0/4.0

BTech in Electrical Engineering, UEM, India — **Aug 2013 – May 2017**
GPA: 7.66/10.0

SKILLS

- **Programming**: Python, SQL, R
- **Data Engineering**: ETL/ELT, Informatica, pySpark, Apache Airflow, Git, CI/CD
- **Cloud & Databases**: GCP BigQuery, Snowflake, PostgreSQL, Teradata, AWS Athena, Amazon S3
- **Modeling & Analytics**: Regression, Classification, Clustering, A/B Testing, Survival and Trend Analysis
- **Data Governance**: Data Validation, Quality Checks, Metadata/Lineage, HIPAA Compliance
- **Domain Expertise**: Healthcare Analytics, Medicaid Claims, EHR Data, OMOP CDM, Epic Data Model
- **Workflow**: Agile, Scrum, Jira

AWARDS AND HONORS

- **Dean's List of Distinguished Graduate Scholars**, University of Arizona iSchool — **Dec 2023**
- **Graduate Research Fellowship with Full Tuition Scholarship**, University of Arizona — **Feb 2023, Aug 2023**
- **Contextual Master Award**, TCS — **Mar 2022**