Gayathri Suresh

Boston, MA 02119 | +1 (857) 263-1831 | suresh.gay@northeastern.edu | linkedin.com/in/gayathri-suresh-063131174

Summary

Data Analyst with a background in electrical engineering and hands-on experience turning large, complex datasets into actionable insights. Skilled in SQL, Python, and Tableau, with a track record of building analytics platforms, statistical A/B testing frameworks, and data quality systems across healthcare, e-commerce, and transportation domains. Experienced with cloud-based data warehouses (Snowflake, BigQuery, Redshift) and dimensional modeling (dbt, star schema, SCD Type 2). Comfortable working across the full analytics stack — from raw data ingestion and transformation to stakeholder-facing dashboards and executive reporting.

Education

Northeastern University Boston, MA
Master of Science in Electrical and Computer Engineering — CV, ML & Algorithms | ***GPA: 3.76*** *Sept 2023 – Dec 2025*
Sri Sivasubramaniya Nadar College of Engineering Chennai, India
Bachelor of Engineering in Electronics and Communication Engineering | ***GPA: 8.84/10.0*** *June 2019 – May 2023*

Technical Skills

Analytics & BI: Tableau, Power BI, Looker Studio, Metabase, KPI dashboards, Business Intelligence reporting
SQL: CTEs, joins, window functions (RANK, PERCENTILE_CONT, LAG/LEAD, STDDEV), data aggregations, subqueries
Data Modeling: dbt (snapshots, tests, marts), star schema, snowflake schema, dimensional modeling, SCD Type 2
Statistical Analysis: A/B testing (chi-square, t-test, Bayesian), hypothesis testing, power analysis, multiple testing correction, regression
Data Warehouses & Databases: Snowflake, BigQuery, Redshift, PostgreSQL, MySQL, MongoDB (NoSQL)
Data Quality & Testing: Great Expectations, Pydantic, dbt tests, schema validation, data profiling, documentation
Data Transformation: dbt transformations, data aggregations, ETL/ELT, data cleaning, normalization
Cloud & Storage: AWS (S3, EC2, Redshift, Glue), GCP (BigQuery), Snowflake, data lakes, lakehouse architecture
Pipeline & Orchestration: Apache Airflow, Dagster, Kafka, Prefect
Programming: Python (Pandas, NumPy, PySpark), SQL, Bash/Shell
Productivity & Reporting: Microsoft Excel (pivot tables, VLOOKUP, charts), GitHub Actions (CI/CD), Docker, Git
Methodologies: Agile, stakeholder management, cross-functional communication, data documentation

Professional Experience

Humatics Corporation Waltham, MA
Associate Information Systems Engineer *June 2024 – Dec 2024; June 2025 – Aug 2025*

- Analyzed real-time railway sensor data across 5+ operational rail corridors, translating time-series signals into safety KPIs and performance reports communicated to operational stakeholders
- Built geospatial data analysis pipeline integrating sensor data, GTFS schedules, and track segment maps using GeoPandas, improving geospatial accuracy of location analytics by 30%
- Integrated 3 external data sources (weather API, GTFS schedules, spatial databases) into the analytics stack, enabling richer predictive models that reduced false safety alerts by 18% — findings presented to cross-functional teams
- Developed data quality validation framework filtering unreliable samples across 77 track segments, with documented schema definitions and validation rules ensuring consistent, trustworthy analytical outputs
- Worked in an Agile environment, participating in sprint planning and iterative delivery of data pipeline features
- Built graph-based network model covering 100% of the track network with <1% error rate, providing a reliable data foundation for downstream safety analysis

Northeastern University Boston, MA
Teaching Assistant — Neural Networks and Deep Learning *Summer 2024*

- Mentored 25–30 students through a rigorous deep learning curriculum, holding weekly office hours and providing one-on-one guidance on ML/deep learning concepts, model implementation, and debugging
- Graded assignments and final projects, providing structured written feedback to strengthen students' understanding of neural architectures and training practices
- Served as primary point of contact for capstone project troubleshooting, communicating complex technical concepts clearly to students at varying skill levels

Datator Information Technology Chennai, India (Remote)
Data Engineering Intern *Oct 2022 – Mar 2023*

- Engineered data collection pipeline using network packet capture tools (PCAPDroid, Wireshark), processing and storing network traffic data in structured CSV format for downstream analysis
- Built ETL workflow extracting 28 time-series and flow-based features from raw network packets using CICFlowMeter, with data transformations and aggregations producing analysis-ready datasets for ML-based security classification
- Implemented preprocessing pipeline (protocol parsing, normalization, clustering) and data quality validation ensuring schema consistency, feature completeness, and packet integrity across multiple sources
- Created comprehensive data documentation including schema definitions, data dictionaries, and pipeline runbooks enabling team collaboration and maintainability

Resileo Labs Chennai, India (Remote)
Data Processing Intern *Jul 2022 – Oct 2022*

- Built medical image data processing pipeline handling CT scans and thermal images, standardizing formats and resolutions to ensure consistent data quality for autoencoder model training
- Implemented data augmentation workflows (rotation, scaling, noise injection, brightness adjustment) expanding training dataset size for improved model generalization
- Designed hierarchical data storage architecture (raw/processed/annotated layers) mirroring a data lakehouse pattern, with metadata tracking system documenting image provenance, processing lineage, and quality metrics

Projects

Healthcare Analytics Platform | *AWS S3, EC2, Snowflake, dbt, Tableau, Power BI, GitHub Actions*

- Analyzed 160K+ CMS Medicare records (inpatient charges, readmissions, hospital quality) using an ELT lakehouse architecture (Bronze/Silver/Gold on AWS S3) with Snowflake as the analytical data warehouse
- Designed a star schema with 4 dimension tables and 4 fact tables enabling self-service analysis of healthcare cost transparency, readmission rates, and hospital quality KPIs across U.S. states
- Wrote 8 dbt mart models and 3 intermediate models with advanced SQL (CTEs, joins, window functions: RANK, PERCENTILE_CONT, LAG/LEAD, STDDEV) for cost benchmarking and outlier detection; documented all models with dbt schema YAML and descriptions
- Built data quality framework achieving 97.7% test pass rate (42/43 dbt tests + 5 Great Expectations suites), flagging 176+ business rule violations without data loss
- Implemented SCD Type 2 via dbt snapshots to track historical changes in hospital ownership, ratings, and services — enabling longitudinal trend analysis
- Created Tableau and Power BI dashboards with interactive KPIs, geographic visualizations, and custom SQL data sources, enabling stakeholder self-service analytics without writing SQL
- Automated CI/CD pipeline with GitHub Actions (dev/prod branch strategy), reducing manual testing effort by 80%+ and preventing invalid data from reaching production

E-Commerce Streaming Analytics Platform | *Kafka, Dagster, BigQuery, Redshift, MongoDB, Python*

- Designed and implemented a statistical A/B testing framework evaluating 10 e-commerce experiments (free shipping, checkout flow, payment installments) using chi-square tests, t-tests, Bayesian analysis, and multiple testing correction
- Produced automated weekly BI reports (HTML/PDF, Excel workbook, executive summary) with experiment recommendations (LAUNCH / ITERATE / CONTINUE) and projected revenue impact for stakeholder communication
- Built real-time event ingestion pipeline processing orders, payments, deliveries, and clickstream data across 7 Kafka topics, achieving end-to-end latency of 2–5 seconds; data landed in both BigQuery (analytical) and Redshift (reporting)
- Designed unified analytics views merging batch (Dagster asset pipeline) and streaming data with deduplication, enabling consistent KPI metrics across historical and live data
- Orchestrated 80+ Dagster assets across batch, streaming, and analysis layers with asset-level data quality checks and documented data lineage blocking downstream materialization on failures
- Implemented multi-store lakehouse architecture (S3 bronze/silver/gold, BigQuery marts, MongoDB NoSQL operational store) supporting ad-hoc analysis and operational reporting

Enterprise Customer Analytics Platform | *Apache Airflow, PostgreSQL, Docker, Metabase, Looker Studio*

- Designed a star schema data warehouse in PostgreSQL with 6 dimension tables and 3 fact tables (transactions, sessions, support interactions) supporting customer behavior and revenue analysis

- Implemented SCD Type 2 for the customer dimension enabling historical customer journey analysis and cohort tracking across 4,268 customer records
- Built aggregated data marts with SQL aggregations producing daily revenue summaries, customer lifetime value (CLV), cohort retention curves, and product performance KPIs for BI consumption in Metabase and Looker Studio
- Developed 4 data transformation modules handling customer deduplication (fuzzy matching via RapidFuzz), multi-currency conversion, and timezone normalization across 5,000 transactions
- Created data quality validation framework (Great Expectations + Pydantic) with automated checks for nulls, duplicates, and business rules; produced data documentation including schema definitions and test coverage reports
- Deployed 3 Metabase dashboards with 100% availability using Docker Compose, with structured logging for pipeline health monitoring