KUMKUM SAREEN

Edison, NJ | (561) 972-1471 | kumkumsareen04@gmail.com | kumkumsareen.me
github.com/kumkumsareen1006 | linkedin.com/in/kumkumsareen

SUMMARY

Computer Science graduate (Rutgers, May 2026) with internship experience validating data across a three-database SQL Server migration. Builds incremental, tested ELT pipelines in Python, SQL, Airflow, dbt, and Snowflake on Docker.

EDUCATION

Rutgers University, School of Arts and Sciences — **New Brunswick, NJ**

Bachelor of Science in Computer Science | GPA: 3.7/4.0 — May 2026

Relevant Coursework: Database Management, Data Structures & Algorithms, Data Science, Statistics, Computer Architecture

PROFESSIONAL EXPERIENCE

EisnerAmper — **June 2025 – August 2025**

Software Engineering Intern — Edison, NJ

- Centralized and validated data from 3 SQL Server databases (50+ tables) during a system migration, designing analytics-ready schemas and enforcing referential integrity across environments.
- Wrote and tuned SQL queries and stored procedures to profile, cleanse, and reconcile legacy data, resolving data quality issues ahead of go-live.
- Built C#/.NET data access with Dapper and standardized object-to-data mappings with AutoMapper, cutting manual transformation errors across the migration path.
- Ran migration testing across source and target systems and documented data lineage for stakeholders, verifying row-level accuracy before production release.

PROJECTS

Vantage — Wikipedia Attention-Spike Lakehouse | PySpark, Delta Lake, Kafka, Structured Streaming, Azure Databricks, ADLS Gen2, Data Factory, pytest

- Built a Bronze/Silver/Gold lakehouse in PySpark and Delta Lake that detects attention spikes in Wikipedia pageviews and live edits, deployed on Azure Databricks with ADLS Gen2 and Data Factory.
- Streamed live edits through Kafka and Structured Streaming; a kill -9 restart test confirmed zero event loss.
- Enforced schemas, quarantined bad rows, and made reruns idempotent with Delta MERGE; pytest parity and restart tests run in GitHub Actions CI.
- Scored spikes with a median/MAD z-score validated against a pandas prototype, separating real news events from bot traffic.

SubscriptionPulse — Incremental ELT pipeline for SaaS subscription analytics

Python, PostgreSQL, AWS S3, Snowflake, dbt, Apache Airflow 3, Docker

- Built an incremental ELT pipeline moving synthetic SaaS data (201 accounts, 19K usage events, 1.5K invoices) from Dockerized Postgres to partitioned Parquet in S3 and into Snowflake via an IAM storage integration.
- Made loads idempotent with updated_at watermarks and COPY INTO load tracking; reruns added zero duplicates and row counts matched at every layer.
- Orchestrated extract, load, and 7 dbt models (including a Type 2 SCD snapshot) as an Airflow 3 DAG, with 32 dbt tests that fail the run on bad data.

E-commerce Analytics Pipeline — Python, PostgreSQL, Pandas, Streamlit

- Built an ETL pipeline ingesting e-commerce data from APIs and flat files into a centralized PostgreSQL database across tables.
- Automated cleaning in Python and Pandas and wrote SQL models analyzing sales, customer behavior, and product trends, surfaced in a Streamlit KPI dashboard.

ADDITIONAL EXPERIENCE

YMCA — **December 2021 – Present**

Client Services Administrator — Edison, NJ

- Processed 200+ data entry tasks in Daxko and cleaned up duplicate member accounts, merging records with different family members and removing exact duplicates.
- Analyzed membership status at the end of each month to track new, churned and paused members, then contacted lapsed members with incentives to re-join.
- Reconciled 500+ member payments and put together membership reports at the end of each season and promotion to measure their effect on enrollment.

SKILLS

Languages: SQL, Python (Pandas), C#, Java

Data & Warehousing: Incremental ELT, Data Pipelines, Dimensional Modeling, Type 2 SCD, Data Quality Testing, Query Optimization

Platforms & Tools: Apache Airflow, dbt, Snowflake, AWS S3, Docker, PostgreSQL, SQL Server (SSMS), .NET, Git, FastAPI, Streamlit