

VIVEK KUMAR

DATA ENGINEER | ETL & REAL-TIME PIPELINE DEVELOPMENT

8092814247 | vivek.kumar281502@gmail.com | [GitHub](#) | [LinkedIn](#) | [HackerRank \(4★ SQL\)](#)

PROFESSIONAL SUMMARY

Data Engineer with hands-on project experience architecting production-style batch and real-time data pipelines that have processed 12,700+ records end-to-end using Python, SQL, Apache Spark, Apache Kafka, and Apache Airflow. Delivers automated, fault-tolerant ETL/ELT systems, star-schema data warehouses, and containerized deployments that cut manual data-handling effort and improve data reliability. GATE 2026 qualified (Data Science & AI) with a strong foundation in data modeling, workflow orchestration, and scalable pipeline design — ready to drive measurable data-infrastructure impact from day one.

CORE SKILLS

- **Languages & Querying:** Python, SQL
- **Data Engineering & Streaming:** Apache Spark, Apache Spark Structured Streaming, Apache Kafka, ETL/ELT, Data Pipeline Design
- **Workflow Orchestration:** Apache Airflow (DAG design & automation)
- **Data Modeling & Warehousing:** PostgreSQL, Star Schema Design, Data Warehousing, Data Validation
- **Libraries:** Pandas, NumPy
- **Cloud & DevOps:** AWS (S3), Docker, Git, GitHub, Containerized Deployment

DATA ENGINEERING PROJECTS

E-Commerce ETL Data Engineering Pipeline — [View Project ↗](#)

- Architected and shipped an end-to-end ETL pipeline (Python, PostgreSQL, Apache Airflow, Docker), automating ingestion and transformation of 11,000+ customer records and 1,700+ sales transactions with zero manual intervention.
- Engineered custom Airflow DAGs to orchestrate scheduled ingestion and transformation jobs, replacing manual data processing with a fully automated, repeatable workflow.
- Built data cleaning, validation, and transformation logic in Pandas to enforce data-quality standards across every record processed.
- Designed a star-schema data warehouse in PostgreSQL, structuring transactional data for fast, analytics-ready reporting.
- Containerized the entire pipeline with Docker, enabling consistent, portable deployment across environments.

Real-Time Order Processing Pipeline — [View Project ↗](#)

- Built a real-time streaming pipeline (Apache Kafka + Spark Structured Streaming + PostgreSQL) to ingest and process live order events with sub-batch latency.
- Implemented in-stream validation, transformation, and revenue-calculation logic, delivering analytics-ready data to PostgreSQL for downstream reporting.
- Designed dedicated error-handling tables to isolate invalid records, preserving data integrity without halting the pipeline.
- Deployed a fault-tolerant, Docker-based architecture engineered for high availability under continuous streaming load.

EDUCATION

B.Tech in Computer Science & Engineering | Sep 2022 – May 2026

Babu Banarasi Das Institute of Technology, Ghaziabad

ACHIEVEMENTS & CERTIFICATIONS

-
- Qualified GATE 2026 (Data Science & Artificial Intelligence) — Score: 349, AIR: 9,935, ranking in the top tier of a national engineering entrance exam.
 - **4★ SQL rating on HackerRank** — [View Profile ↗](#)