Asparsh Raj

+91-7667406415 | rajasparsh19@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Data Engineer with a proven track record of building **core data products** and establishing a **single source of truth** for massive datasets. Specialized in architecting high-performance RAG systems, managing **1.65 Crore+ records**, and optimizing cloud infrastructure costs by **75%**. Experienced in distributed systems, leveraging Python, Airflow, SQL and DBT to democratize data insights and drive business growth.

EDUCATION

Sarala Birla University **Sept, 2022 – Aug, 2025**
Bachelor of Computer Applications (BCA) - ***CGPA: 8.9*** *Ranchi, India*

TECHNICAL SKILLS

Data Engineering: Airflow, DBT (Core), ETL/ELT, VM Orchestration, Web Scraping (Playwright, Selectolax)
Cloud & Infra: AWS (S3, Lambda, API Gateway), GCP (Google Cloud Storage), Databricks, Linux, SSH
Databases: Postgres, MongoDB, Meilisearch
Languages/Frameworks: Python, SQL, FastAPI, Django, DBT Macros

EXPERIENCE

Claw Legaltech Pvt Ltd **April, 2025 – Present**
Data Engineer (Full-time) *Ahmedabad, Gujarat*

- **Single Source of Truth:** Established a centralized repository by migrating complex legal datasets from Azure AI Search to a custom **Postgres (pgvector)** solution, reducing costs by **75%** and ensuring 100% data consistency.
- **Large-Scale Data Assembly:** Engineered pipelines to structure **1.65 Crore (16.5M) judicial records**, optimizing query performance from 4 minutes to **3 seconds** via HNSW indexing and state-based partitioning.
- **Core Data Products:** Built a mission-critical search engine tracking 814 district courts using **20+ parallel AWS Spot VMs** and S3-based state checkpoints, securing Tier-1 enterprise clients like **HDFC and RPSG Group**.
- **Complex Ingestion:** Developed a robust ETL framework using **Airflow and FastAPI** middleware, handling **5,000+ daily live-data calls** and automating extraction across 40+ high-protection legal portals.
- **Data Pipeline Scalability:** Optimized **Python pipelines** and deployed on Airflow as DAGS, improving overall ETL efficiency by reducing manual intervention and supporting rapid organizational scaling.

Claw Legaltech Pvt Ltd **Dec, 2024 – March, 2025**
Data Engineer Intern *Ahmedabad, Gujarat*

- **Democratizing Insights:** Built modular **DBT models** (Staging, Intermediate, Mart) and implemented **DBT Tests**, ensuring high data quality and reliability for downstream business analytics users.
- **Efficiency Engineering:** Reduced lines of code by **95%** and development time by **78%** through the implementation of **DBT Macros** and dry-coding standards, doubling team output.
- **Automated Workflows:** Orchestrated end-to-end production pipelines on **Databricks Cloud** using **Apache Airflow**, maintaining high availability for real-time reporting.

PROJECTS

Legal RAG Retrieval Engine | *Postgres, pgvector, Airflow, GCP*

- Assembled a high-performance vector search engine for 16.5M documents, enabling semantic search and democratizing access to legal precedents across all Indian High Courts and the Supreme Court.
- Automated the end-to-end data lifecycle in Apache Airflow, integrating PDF extraction, embedding generation, and record upserts to eliminate manual case updation efforts.

Distributed Captcha Solver Service | *FastAPI, Python*

- Engineered an in-house ML-based middleware to bypass complex captchas, reducing external API dependencies and costs to zero while serving 5,000+ stateless requests daily.
- Developed an extensible OOP framework allowing for the seamless integration of new solver modules for evolving captcha patterns.

CERTIFICATIONS

PySpark - Apache Spark Programming in Python - Udemy