VEDANT ACHOLE

8578328355 | vedant4815@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Data Engineer and AI practitioner with 1+ year of professional experience delivering cloud-based data solutions and production-grade AI systems. Built a healthcare payment integrity platform on Databricks/AWS processing $500B+ in Medicare claims with XGBoost fraud detection and SHAP explainability. Designed end-to-end ELT pipelines using PySpark, dbt, and AWS Glue on medallion architecture, and engineered an LLM-powered RAG system using OpenAI GPT and FAISS. Skilled across the full data engineering lifecycle - ingestion, transformation, orchestration, quality, and serving - with hands-on experience in Python, SQL, Spark, and both AWS and Azure cloud stacks. Collaborative Agile team contributor with proven ability to translate business requirements into scalable, production-ready data and AI solutions.

TECHNICAL SKILLS

- **Programming & Data**: Python, PySpark, Pandas, NumPy, SQL, REST APIs, JSON
- **Cloud & Data Engineering**: AWS, S3, Glue, Lambda, Athena, EMR, Step Functions, IAM, CloudWatch, Azure, Databricks, Data Factory, dbt, Airflow, ELT/ETL, Medallion architecture, Parquet/Delta Lake, Snowflake
- **AI & Machine Learning**: SHAP, Scikit-learn, LLMs, OpenAI GPT, Hugging Face Transformers, RAG, FAISS Vector DB, Prompt Engineering, NLP
- **DevOps & Tools**: Git, GitHub, CI/CD workflows, Docker, FastAPI, Streamlit, JIRA, Tableau, MS Excel

KEY PROJECTS

Healthcare Payment Integrity & Fraud Detection Platform | link **Apr 2026**

BillingShield

- Stack: Databricks , dbt, PySpark, XGBoost, SHAP, FastAPI, Streamlit · Data: CMS Medicare (~$500B claims universe)
- Engineered end-to-end ELT pipeline on CMS Medicare data using PySpark and dbt, processing raw claims into curated Bronze/Silver/Gold Lakehouse layers with schema validation, referential integrity checks, and Delta table format optimization.
- Trained XGBoost gradient boosting classifier for fraud and anomaly detection across provider billing patterns; applied SHAP explainability layer to surface human-readable risk rationale for each prediction, enabling non-technical stakeholder consumption.
- Implemented Dagster orchestration with dependency management, retry logic, and scheduling - mirroring production Airflow/MWAA workflow patterns for reliable pipeline execution.
- Deployed FastAPI inference API and Streamlit risk dashboard enabling analysts to self-serve fraud scores, drill into provider cohorts, and export risk tier summaries - reducing engineering dependency for common data access.
- Identified $2.3B+ in anomalous billing patterns across provider cohorts; built Critical/Elevated/Normal risk stratification tiers across 500B+ records, demonstrating end-to-end ML pipeline delivery from raw ingestion to production serving.

Healthcare Claims Analytics Pipeline | link **Mar 2026**

- Stack: AWS Glue, PySpark, S3, Athena · Medallion architecture (Bronze → Silver → Gold)
- Built production ELT pipeline using AWS Glue (PySpark) across 4 normalized claims tables - joining data, applying window functions for provider rankings, computing derived fields, and delivering curated Gold analytics datasets in Parquet columnar format.
- Delivered 6 Gold-layer financial KPIs: high-cost claims (top 1%), fraud scoring (229 flagged claims, $25M+), provider performance rankings, regional exposure analysis, diagnosis cost breakdown, and patient risk stratification (193 critical cases).
- Automated schema detection and data cataloging via AWS Glue Crawlers; cataloged 10 tables in Glue Data Catalog with Parquet optimization achieving sub-second Athena query performance on large-scale claims data.
- Quantified business insights: identified $88.9M in cancer treatment costs, detected $25M+ in fraud patterns, and prioritized high-risk patient cohorts - demonstrating ability to translate raw financial data into actionable analytics.

LLM-Powered Resume Matching System

AI Talent Marketplace

- Stack: Python, OpenAI GPT-4, Hugging Face Transformers, FAISS Vector DB, Sentence Transformers, RAG
- Built end-to-end RAG pipeline combining dense vector search (FAISS) with GPT-4 re-ranking to semantically match 3,000 candidate profiles against open roles - overcoming limitations of keyword-based ATS systems.
- Demonstrated live proof-of-concept to CGI Technology senior leadership achieving 40% improvement in candidate relevance over baseline keyword matching - showing ability to communicate complex AI outputs clearly to executive audiences.

PROFESSIONAL EXPERIENCE

CGI Technology **Sep 2025 - Present**

Consultant - Data Validation & Analytics *Lafayette, LA*

- Execute 50+ business scenarios and ~1,200 test cases per release cycle; maintain structured defect documentation and workflow tracking in JIRA, preventing post-release data quality issues.
- Demonstrated internal AI proof-of-concept (LLM resume matching system) to CGI Technology senior leadership, achieving 40% improvement in candidate relevance and bridging ML engineering with clear business communication.
- Collaborate in Agile squads with developers, business analysts, and QA leads to clarify data requirements and confirm acceptance criteria under sprint deadlines, enabling on-time delivery of features

KPMG | BSR & Co. LLP **Feb 2024 - Aug 2024**

Analyst (Internship) *Pune, India*

- Performed systematic data validation comparing frontend BI dashboard outputs against backend database tables using SQL, confirming financial data accuracy for regional managers across vehicle sales and inventory reporting systems.
- Identified and documented a critical cross-module data persistence defect affecting financial reporting accuracy; prepared structured reproduction steps enabling swift remediation by the development team.

EDUCATION

Boston University – Questrom School of Business **May 2025**

Master of Science, Management *Boston, MA*

- **Achievements:** Director's Honors
- **Coursework:** Agile Project Management, Data Management for Managers, Quantitative Business Analysis

Vishwakarma Institute of Technology **May 2024**

Bachelor of Technology, Artificial Intelligence & Data Science *Pune, India*

- **Coursework:** Machine Learning, Data Structures & Algorithms, Database Management Systems, Big Data Analytics, Statistical Methods