VIKAS RATHI

rathivikas06@gmail.com ◇ 6375334007 ◇ linkedin.com

SKILLS

Core:	Spark, Python, SQL, Airflow, PySpark
Cloud Platforms:	Azure (Fabric, Databricks, Data Factory, Synapse), GCP (BigQuery, Composer), AWS (Redshift, Lambda)
AI & GenAI:	LangGraph, Prompt Engineering, Google ADK
DevOps & Tools:	Docker, Git
Core Concepts:	System Design (LLD/HLD), Distributed Systems, Agile

WORK EXPERIENCE

Software Development Engineer Sept 2024 – Present

Sigmoid Analytics

- Architected a multi-agent Text-to-SQL system that improved query accuracy from 42% → 85%, reducing manual SQL curation effort by 60% and accelerating analytics turnaround time for business users.
- Designed and implemented an Azure Databricks–based Medallion Architecture handling 500GB+ daily data, leveraging partitioning, Delta optimizations, and SCD Type-2 modeling to reduce Power BI query latency by 65%.
- Designed and orchestrated end-to-end data workflows with Azure Data Factory, managing inter-job dependencies across distributed Azure Databricks pipelines and implementing automated retries, alerting, and failure handling to reduce data downtime and operational incidents.
- Designed CoT-driven prompting that lowered hallucinations from 38% to 14%, significantly decreasing invalid SQL generation and stabilizing agent behavior across complex business queries.

Data Engineer Oct 2021 – Sept 2024

Accenture

- Built ingestion workflows as part of a 1TB on-prem to Google Cloud migration, handling both historical datasets and 40-50GB of daily incremental loads using Cloud Composer, Dataproc, and BigQuery.
- Buiked a configuration-driven Airflow framework that decoupled ingestion logic from parameters, enabling the team to scale to 800+ DAGs with zero code duplication and significantly reduced maintenance overhead.
- Developed pre-ingestion data quality checks and a synthetic data creator to validate schema, formats, and edge cases before loading data into GCS, improving pipeline reliability.
- Designed and deployed Python-based OCR parsing services on Kubernetes to process large unstructured healthcare claim PDFs (1000+ pages), extracting dates, diagnoses, and cost details and persisting structured outputs into Azure Databricks Delta tables to support 500K reimbursement calculations at scale
- Applied GDPR/HIPAA-aligned practices such as masking, restricted access, and audit logging while handling sensitive healthcare and financial data.

TECHNICAL PROJECTS

- **CDC Pipeline with Spark Structured Streaming:** Designed a low-latency Change Data Capture (CDC) pipeline ingesting transactional data into Delta Lake using Kafka and Spark Structured Streaming.
- **MCP Server for Databricks SQL Queries:** Built an MCP server allowing LLMs to safely execute SQL queries on Databricks, with schema discovery, query validation, secure authentication and LLM-friendly response formatting.

EDUCATION

Bachelor of Technology in Computer Science, Rajasthan Technical University 2017 – 2021

CERTIFICATIONS

- **Software Development and Problem Solving**, Scalar 2023 – 2024
  Coursework: Data Structures and Algorithms, Low-Level Design, High-Level Design
- Google Associate Data Engineer (GCP)
- Microsoft Certified: Fabric Analytics Engineer Associate
- Microsoft Certified: Azure AI Engineer Associate (AI-102)
- Databricks Certified Associate Data Engineer

LEADERSHIP & ACHIEVEMENTS

- Winner of Company-Wide Hackathon (Sigmoid Analytics): Built a no-code, config-driven ETL accelerator on Microsoft Fabric that enables automated pipeline creation using Fabric notebooks for ingestion and validation.
- Mentored 8 interns in designing a Medallion Architecture–based data pipeline and building a Streamlit dashboard for ad-driven revenue metrics.