Rahul Chavan

rcchavan663@gmail.com | GitHub | LinkedIn

Pune, Maharashtra
Mobile: +91-85549-38009

Data Engineer

Results-driven Data Engineer with **2+ years** of experience building scalable ETL pipelines and real-time data solutions across AWS, Azure, and Databricks. Skilled in Python, SQL, and PySpark with a strong focus on SAS-to-PySpark migrations, Generative AI integration, and modern cloud data architectures.

Technical Skills

Cloud Platforms	:	Google Cloud Platform (GCP), Amazon Web Services (AWS), Microsoft Azure
Data Warehouse	:	Snowflake, AWS Redshift
ETL Tools	:	AWS Glue, AWS Athena, Azure Data Factory (ADF)
Data Platforms	:	Azure Databricks, Microsoft Fabric
Languages	:	Python, SQL, PySpark, SAS
BI Tools	:	Power BI, AWS QuickSight
AI / ML	:	Large Language Models (LLMs), RAG, Generative AI

Experience

LTIMindtree Sep 2023 – Present
Data Engineer *Remote*

- Built GenAI-powered applications for tasks like code translation, lineage extraction, and metadata classification by integrating LangChain, OpenAI APIs, Vertex AI, and AWS SageMaker into robust backend pipelines. These apps enhanced automation and reduced manual engineering by 60%.
- Enabled natural language to SQL and PySpark translation within Streamlit dashboards, allowing business users to generate and execute ad-hoc queries across cloud datasets without engineering involvement. Added support for schema introspection, auto-correction, and result previews.
- Designed and implemented intelligent agent-based systems—Meta Migrator, Analyzer, and Validator—for automating SQL-to-PySpark code conversions. These agents collaborated using task-specific prompt chains, validation heuristics, and rule-based branching.
- Built a multi-agent orchestration framework with caching, retry logic, schema enforcement, transformation-type detection, and logging to ensure explainable, traceable, and accurate code conversions in regulated environments.
- Developed modular CI/CD pipelines for PySpark and GenAI workflows using GitHub Actions (automated testing and deployment), Docker (containerization), and Terraform (cloud infrastructure provisioning), ensuring reproducible and scalable deployments across AWS and Azure.
- Optimized Databricks performance by tuning Spark configurations, applying job-level parallelization strategies, enabling autoscaling clusters, and eliminating I/O bottlenecks, resulting in an 18% reduction in compute cost and faster execution of large ETL jobs.
- Integrated enterprise-grade data pipelines with Snowflake, S3, Delta Lake, and Azure Data Lake, enabling secure and scalable movement of structured and semi-structured data across cloud platforms in alignment with enterprise governance standards.
- Implemented robust data quality validation layers using Great Expectations and custom PySpark UDFs to enforce schema consistency, detect anomalies, and ensure accurate, high-integrity data processing across batch and streaming pipelines.
- Currently working on multi-agentic Retrieval-Augmented Generation (RAG) pipelines and LLM fine-tuning for data migration use cases across SAS, PL/SQL, and legacy enterprise workloads.
- Focused on enhancing transformation accuracy by combining vector-based search, prompt engineering, and domain-specific model tuning to automate schema mapping, lineage extraction, and SQL-to-PySpark conversion.
- Currently researching and building multi-agentic RAG systems with fine-tuned LLMs to automate schema mapping, metadata enrichment, and transformation logic detection for large-scale enterprise data migrations.
- Built scalable data pipelines on Microsoft Fabric using Lakehouse architecture and Dataflows Gen2, enabling unified batch and streaming ETL across OneLake, Synapse, and Power BI.
- Integrated Power BI and OneLake for real-time reporting over curated Delta tables in Fabric Lakehouse.

Projects

Pik-Pok AWS EC2, Lambda, S3, CloudFront
Serverless Social Video Platform

- Built a scalable social video platform using AWS Lambda and S3 for content storage and delivery, integrated with CloudFront for low-latency streaming.
- Implemented WordPress backend for CMS and configured IAM roles and policies for secure media access.
- Enabled real-time engagement features like likes, shares, and comments with API Gateway and DynamoDB.

MediBot Streamlit, Google Gemini API
AI-Powered Health Assistant

- Developed an interactive Streamlit application that answers patient queries using Google Gemini LLM APIs.
- Trained on hospital SOPs and FAQs to provide consistent and accurate medical guidance.
- Enabled multilingual support and dynamic prompt templating for personalized AI responses.

Text-to-SQL Web App Streamlit, Gemini, TogetherAI
Natural Language to SQL Generator

- Built a web-based application that transforms natural language queries into optimized SQL queries using LLMs from Gemini and TogetherAI.
- Enabled XLSX upload, schema parsing, and prompt-based inference for data-specific SQL generation.
- Improved user experience with error-handling logic, auto-completion, and explainable query outputs.

SAS to PySpark + LLM Automation Databricks, Streamlit, Python, AWS, GCP
Intelligent ETL Modernization Platform

- Migrated SAS workflows to PySpark using Databricks notebooks and Python-based analyzers, enabling automated lineage and dependency extraction.
- Developed regex-based parsers to classify metadata and detect transformation patterns for accurate code conversion.
- Integrated SageMaker and Vertex AI for LLM fine-tuning and prompt engineering for text-to-SQL and text-to-PySpark translation.
- Built Streamlit-based dashboards for executing and validating LLM-generated code using interactive interfaces.

Modern Data Lake Architecture AWS Glue, S3, Athena, Redshift, Power BI
Serverless Data Lake + Analytics Platform

- Designed and built a serverless data lake on AWS using Glue for ETL, S3 for storage, and Athena for ad-hoc querying.
- Automated metadata cataloging with Glue Data Catalog and optimized transformation jobs with PySpark.
- Integrated Redshift with S3 for seamless data warehousing and analytical processing.
- Developed Power BI dashboards on top of Redshift and Athena to visualize operational metrics.
- Enabled basic ML capabilities using AWS SageMaker, Comprehend, and Rekognition for enriched analytics.

Certifications

- AWS Data Engineer Associate
- Snowflake SnowPro Core
- Databricks Data Engineer Associate
- Databricks Generative AI Engineer Associate
- Oracle Generative AI Certified
- Microsoft Fabric Analytics Associate

Education

B.Tech in Civil Engineering Jul 2018 – Aug 2022
SKN Sinhgad College of Engineering *CGPA: 9.8*

HSC (Science) 2018
Maharashtra State Board *Percentage: 77.78%*