YADUKRISHNAN J

Bengaluru, India | krishnanyadu60@gmail.com | +91 8660042893 | linkedin.com/in/yk9238 | github.com/Yadu9238 | yadu.dev

PROFESSIONAL SUMMARY

Data Engineer with 4 years building distributed pipelines and lakehouse platforms on Databricks, PySpark, and SQL, plus production retrieval-augmented generation systems on Azure OpenAI and Azure AI Search. Engineered a metadata-driven ingestion framework processing 50+ TB and 4M+ objects at ~850 GB/day with automated reconciliation, unlocking up to $1.5M in annual cost avoidance. Strong in Spark tuning, SQL optimisation, dimensional and medallion modelling, and CI/CD for production reliability.

TECHNICAL SKILLS

Languages & Processing: Python, SQL (advanced), PySpark, Apache Spark, Spark SQL

Data Platform: Azure Databricks, Delta Lake, Databricks Workflows, ADLS Gen2, Azure Data Factory, Azure SQL Database, Key Vault, Azure AD

Data Engineering: Distributed batch pipeline design, ETL/ELT, orchestration, Medallion Architecture, dimensional modelling, star schema, SCD Type 1/2, partitioning and performance tuning, data quality and reconciliation frameworks

AI & Retrieval: RAG, Azure OpenAI, Azure AI Foundry, Azure AI Search, embeddings, chunking strategies, vector and semantic retrieval

Practices & Tools: Git, GitHub Actions, Azure DevOps, CI/CD, SonarQube, Apiiro, REST and Graph APIs, Agile Scrum, Power BI; working knowledge of Kafka streaming design and Airflow DAGs

Certification: Databricks Certified Data Engineer Associate

PROFESSIONAL EXPERIENCE

Data Engineer | Shell Nov 2025 - Present

- Architected a retrieval-augmented generation assistant over three years of incident records and 100+ technical manuals, replacing a full-context-window approach with vector retrieval on Azure AI Search and Azure OpenAI to ground answers in structured and unstructured enterprise data.
- Engineered the retrieval layer end to end - parsing, chunking, embeddings, semantic search, and prompt grounding - tuning chunk size and retrieval depth against a curated question set to lift answer relevance and cut tokens per query; shipped to pilot with Azure AD-gated access and GitHub Actions CI/CD.
- Designed a metadata-driven ingestion framework in Python, PySpark, and Azure Data Factory pulling from REST and Graph APIs into a Delta lakehouse, processing 50+ TB and 4M+ objects via schema-driven configuration instead of per-source code, unlocking up to $1.5M in annual cost avoidance.
- Tuned Spark and ADF pipelines for parallelism, partitioning, and API throttling to sustain ~850 GB/day across concurrent pipelines each processing 200K-300K+ records per run, with checkpointing and idempotent retries for safe reprocessing.
- Built bronze/silver/gold pipelines on ADLS Gen2 and Databricks with SCD Type 1/2 dimensional models, plus an automated reconciliation layer comparing source and target at record, volume, and checksum level that made acceptance criteria measurable and enabled zero-rollback releases.
- Technical lead for a five-member team; gathered requirements from on-site Upstream asset focals, architects, and product owners, enforced quality and security gates via GitHub Actions with SonarQube and Apiiro, and defined data management standards adopted across global business units.

Associate Data Engineer | Shell Aug 2022 - Nov 2025

- Designed star-schema models and the semantic layer for Upstream asset performance and operations analytics, defining grain, hierarchies, and business rules with on-site asset focals.
- Tuned SQL queries, indexing, and warehouse compute on high-volume analytical workloads, cutting average dashboard load time by 35% and lowering compute consumption.
- Built Python and SQL pipelines onboarding enterprise datasets to Azure across cleansing, business-rule application, and validation, and resolved 200+ production incidents within SLA, eliminating recurring failure classes with durable fixes and automated alerting.

Machine Learning Engineer Intern | ABB Jan 2022 - Jul 2022

- Built Python ETL pipelines and trained a material-strength prediction model, deploying it on an Azure pipeline with automated ingestion and scheduled retraining.

EDUCATION

Ramaiah Institute of Technology | B.E., Information Science and Engineering | CGPA 9.28 / 10 Aug 2018 - Jul 2022