

Don Richardson Bayya

Dallas, TX (Open to Relocate) | +1-857-397-2362 | donrichardsonbayya@outlook.com | [linkedin.com/in/donrichardsonbayya1](https://www.linkedin.com/in/donrichardsonbayya1) | github.com/donrichardsonbayya | [Portfolio](#)

Summary

Data Engineer with strong experience designing and optimizing **scalable ETL and ELT pipelines**, **data warehouse** and **Lakehouse architecture**, and cloud-based analytics systems. Skilled in building **modern data stacks** across **Azure, AWS, and GCP**, leveraging **Databricks, Snowflake, and Redshift** for high-performance data engineering. Proficient in developing **batch and streaming data pipelines** using **Apache Beam, PySpark, and Airflow** to enable real-time insights and event-driven analytics. Proven ability to **automate ingestion, transformation, and deployment workflows** through **CI/CD pipelines** in **Azure DevOps and GitHub Actions**, ensuring data reliability and scalability. Experienced in implementing **data governance, semantic/analytics layers**, and **cloud-native, cost-optimized architectures** to reduce compute expenses and accelerate business decision-making.

Technical Skills

- **Programming Languages:** Python, SQL, PySpark, DAX
- **Data Engineering & Orchestration:** Databricks, Azure Data Factory, AWS Glue, Apache Spark, Apache Beam, Airflow, dbt, Dagster, Airbyte
- **Cloud & Infrastructure:** Azure, AWS, Google Cloud Platform, Hadoop, Hive, Snowflake, CI/CD, Infrastructure as Code (IaC), Cloud-Native Architecture
- **Databases & Warehouses:** PostgreSQL, MySQL, Snowflake, Redshift, BigQuery, MongoDB
- **Data Modeling & Architecture:** Dimensional Modeling, Star Schema, Data Warehouse / Lakehouse Design, ETL/ELT Pipelines
- **Analytics & Visualization:** Power BI, Tableau, Grafana
- **DevOps & Containers:** Docker, Docker Compose, Azure DevOps, Git, GitHub Actions
- **Machine Learning & Libraries:** scikit-learn, TensorFlow, XGBoost, Keras, NLP
- **Tools & Governance:** REST APIs, Data Quality Testing, Monitoring, Orchestration, Data Governance

Professional Experience

Big Data Engineer **Celebal Technologies** on Futureense Technologies' payroll India | 2022–2024

Beverages Company | Azure Data Factory & Databricks ETL

- **Designed and deployed 15+ end-to-end ETL pipelines in Azure Data Factory and Databricks** to centralize market and retail data ingestion from multiple **APIs and flat-file sources**.
- **Engineered** parameterized pipelines for modular data ingestion, enabling environment-based configurations and **CI/CD deployment** across dev, test, and prod environments.
- **Optimized Spark cluster** resource allocation and partition strategies, resulting in a **35% reduction in compute costs and faster data processing cycles**.
- **Integrated Azure Logic Apps** for dynamic failure notifications and escalation workflows, **reducing reporting delays by 80%** and improving SLA adherence.
- **Introduced Grafana dashboards** for real-time job health monitoring, bringing full observability to ETL operations and **reducing troubleshooting time by 50%**.

***Tech Stack:** Azure Data Factory, Databricks, PySpark, Logic Apps, Grafana, Azure SQL*

E-commerce Sector | Unity Catalog Migration

- **Led the end-to-end migration of 200+ Hive Metastore tables, views, and workflows into Databricks Unity Catalog**, ensuring **schema consistency, data governance, and secure access control** across multiple environments.
- **Automated** the entire migration process through **Databricks Notebooks and REST APIs**, **cutting manual workload by 60%** and accelerating overall migration timelines.

Don Richardson Bayya

Dallas, TX (Open to Relocate) | +1-857-397-2362 | donrichardsonbayya@outlook.com | [linkedin.com/in/donrichardsonbayya1](https://www.linkedin.com/in/donrichardsonbayya1) | github.com/donrichardsonbayya | [Portfolio](#)

- **Refactored and validated 50+ ETL jobs** for Unity Catalog compatibility, resolving data access inconsistencies and reducing ETL failure rates by 30%.
- Established a standardized **metadata governance layer**, improving catalog discoverability across production and development environments.
- **Delivered performance improvements** in job orchestration and dependency handling **through optimized cluster configurations and Spark tuning**.

Tech Stack: Databricks, PySpark, Unity Catalog, Azure Data Factory, REST API, SQL

Food Company | AWS Glue Data Platform

- **Architected a scalable AWS Glue data platform** to consolidate data from **SAP systems, S3, and on-prem data sources** into **Amazon Redshift, enabling near-real-time analytics**.
- Refactored legacy **Python ETL scripts** into Glue workflows, introducing parameterized job triggers and parallel task execution **reducing reporting turnaround by 40%**.
- **Optimized Spark execution plans** through predicate pushdown, partition pruning, and broadcast joins, **improving query performance by 25%**.
- **Implemented** incremental load management using Glue bookmarks and S3 partitioning, **ensuring data freshness and minimal reprocessing**.
- **Delivered a modular pipeline framework** now used for multi-region ingestion, improving maintainability and scaling efficiency.

Tech Stack: AWS Glue, Redshift, S3, Spark, Python, SQL, Boto3

Snowflake to Databricks Migration

- **Automated the conversion and migration of 500+ Snowflake SQL scripts**, stored procedures, and analytical workflows **into Databricks SQL and PySpark environments**.
- Developed Python-based lineage tracking scripts to analyze dependencies, object references, and transformation logic prior to conversion.
- Created conversion accelerators for DDL, DML, and procedural logic (e.g., stored procedures and UDFs), **resulting in 50% faster migration execution**.
- **Designed and deployed validation scripts** comparing pre-migration and post-migration **data quality and performance metrics, ensuring accuracy and reliability**.
- Delivered a reusable migration framework now standardized across three enterprise-scale data modernization programs, **cutting future migration effort by over 40%**.

Tech Stack: Snowflake, Databricks, Python, SQL, Azure DevOps, Git, API Automation

Data Specialist Trainee

Futureense Technologies India | 2021–2022

- **Wrote and optimized SQL for extraction**, cleaning, and transformation from multiple sources.
- **Exposure to file formats Avro, Parquet and Lakehouse table formats (Apache Iceberg)**; used Parquet for columnar storage/partitioning and handled basic schema evolution.

Academic Projects

PeeringDB Analytics Pipeline (End-to-End DE + Analytics Project) | [GitHub](#)

- **Developed a containerized data pipeline** integrating PeeringDB API with **Airflow, PostgreSQL, dbt, and Power BI** for telecom infrastructure **analytics**.
- **Automated 6-hour refresh cycles**, delivering real-time **utilization dashboards revealing 90% underutilization** of network capacity and geographic concentration insights.

Tech Stack: Airflow, dbt, PostgreSQL, Power BI, Docker, Python

Don Richardson Bayya

Dallas, TX (Open to Relocate) | +1-857-397-2362 | donrichardsonbayya@outlook.com |
[linkedin.com/in/donrichardsonbayya1](https://www.linkedin.com/in/donrichardsonbayya1) | github.com/donrichardsonbayya | [Portfolio](#)

Real-Time Order Analytics Pipeline (Streaming DE Project) | [GitHub](#)

- **Built a real-time data pipeline** simulating continuous e-commerce orders processed through **Apache Beam** and stored in **PostgreSQL**.
- **Implemented 1-minute tumbling window aggregations to compute revenue** and order counts, **containerized deployment** for reproducible local runs and BI consumption.

Tech Stack: Apache Beam, PostgreSQL, Python, Docker

Modern ELT Stack (PostgreSQL → Airbyte → BigQuery → dbt → Dagster) | [GitHub](#)

- **Implemented a cloud-native ELT pipeline** automating extraction and transformation for e-commerce data.
- **Integrated Airbyte with BigQuery and dbt transformations orchestrated by Dagster**, achieving **95% reduction in manual data handling and 24/7 data freshness**.

Tech Stack: Airbyte, Dagster, dbt, BigQuery, Python, SQL

Power BI Analytics Dashboards | [GitHub](#)

- **Designed and developed interactive Power BI dashboards** covering both global well-being trends and healthcare claims utilization, integrating insights across social and financial datasets.
- **Analyzed** nine years of World Happiness data alongside healthcare cost drivers, **implementing** PMPM, **correlation** metrics, and time-series visuals to uncover patterns in GDP, health, and claim expenditure.
- **Delivered KPI cards**, cost deviation tracking, and comparative **analytics** that improved data transparency and executive decision-making for cross-domain insights.

Tech Stack: Power BI, MySQL, DAX, Power Query, SQL

Education

Hult International Business School — Boston, MA

Master of Science in Business Analytics, 2025

Chaitanya Bharathi Institute of Technology — Hyderabad, India

Bachelor of Engineering in Information Technology, 2021

Certifications

- Data Engineering Hands-On Practice – LinkedIn Learning
- Google Advanced Data Analytics – Google (Coursera)
- Advanced MySQL Topics – Meta (Coursera)

Certifications In Progress:

- AWS Certified Data Engineer-Associate
- Microsoft Certified: Fabric Analytics Engineer Associate

Core Strengths

- End-to-end **data pipeline** development across batch, streaming, and **ELT workflows**
- Strong grasp of **data modeling, orchestration, and transformation** using **Databricks, dbt, and Airflow**
- **Hands-on experience with Azure, AWS, and GCP ecosystems**
- **Automation-first** mindset: optimized costs, reduced latency, and **improved data reliability**
- **BI and analytics** integration for real-time business decision support
- **Expertise in SQL optimization, Spark tuning, and pipeline observability**
- **Proven capability** to align **engineering with business value** through **data-driven outcomes**