

HARISH RAO YADAGIRI

Austin, TX | 737-420-1959 | harishrao@usf.edu | linkedin.com/in/harishrao0724 | github.com/harishraoyadagiri

PROFESSIONAL SUMMARY

Data Engineer with 4+ years building production-grade ETL pipelines, AI-driven automation, and BI/DW platforms across enterprise environments. At Wipro, I used PySpark to process large-scale Jira log data and built Azure Data Factory workflows that automated multi-stage data movement across cloud systems. At Tiruven, I moved into AI-powered data architecture — designing canonical data models, deploying autonomous AI agents, and building the data foundation that executives rely on for emissions and efficiency analytics. I hold a Databricks Generative AI certification and bring hands-on experience with the modern data stack: Spark, Python, Databricks, AWS, and REST API integration.

WHAT I WORK WITH

HOW I SPEND MY TIME

ETL & Pipeline Engineering End-to-end pipeline design using PySpark and Python — from raw ingestion to transformation and loading. At Wipro, this meant processing high-volume Jira ticket logs and structured enterprise datasets; at Tiruven, it means managing emissions data flows that feed CEO-level dashboards.

Workflow Orchestration Used Azure Data Factory to automate multi-stage data movement and transformation across cloud environments at Wipro — scheduling, dependency management, and error handling across 15+ reusable pipeline frameworks.

AI-Driven Data Workflows Built an autonomous AI agent at Tiruven that uses synthetic data simulation to predict employee efficiency drivers. Also applied NLP and Transformer-based models (spaCy, Hugging Face) in a chatbot project with 89% intent recognition accuracy — end-to-end, from training to Flask/FastAPI deployment.

Canonical Data Modeling & Entity Resolution Designing schemas where a 'Customer' in Sales means the same thing as a 'Customer' in Product and Marketing. At Tiruven, this is the core of what I do — creating unified data models across 10+ enterprise domains so stakeholders see one consistent version of the truth.

Data Quality & Normalization Automated cleaning, deduplication, and enrichment workflows that run without manual intervention. Resolved 50+ high-severity defects in production pipelines at Wipro through systematic root-cause analysis.

Python Service Development & REST API Integration Built Python data services deployed via Flask and FastAPI; integrated with external REST APIs including authentication flows, rate limiting, and retry logic — production-hardened, not just prototype-grade.

CI/CD & Infrastructure Automation Containerized environments with Docker; infrastructure provisioning via Ansible and Bash. At Wipro, this cut environment setup time by 40% and eliminated configuration drift across virtualized compute environments.

TOOLS & TECHNOLOGIES

Python Primary language across all roles — PySpark (log processing, large-scale ETL), pandas & NumPy (data transformation), spaCy & Transformers (NLP/ML), Flask & FastAPI (service deployment).

Apache Spark / PySpark Used at Wipro to process large-scale Jira log datasets and enterprise-wide structured data pipelines — partitioning, optimization, and distributed transformation at volume.

Databricks Certified (Generative AI Fundamentals). Hands-on with the Databricks ecosystem for ML model development, notebook-based workflows, and Delta Lake-compatible data architecture.

Azure Data Factory Orchestrated multi-stage ETL workflows at Wipro — data movement, scheduling, trigger-based pipeline execution across cloud and on-prem sources.

AWS Cloud infrastructure for compute, storage, and data services — used across pipeline deployments and BI platform hosting.

SQL (MySQL, Oracle SQL, MongoDB) Complex analytical queries, schema design, and database-backed survey automation; SQL was the foundation of dashboard work at Amrita and data validation frameworks at Wipro.

BI & Visualization Tableau and Power BI for executive-facing dashboards (including the CEO Insights platform at Tiruven); Looker Studio and Qlik also in toolkit.

DevOps & Infrastructure Docker (containerization), Ansible (configuration management), Bash scripting, Git/GitLab — used together for CI/CD pipelines and repeatable infrastructure provisioning.

PROFESSIONAL EXPERIENCE

Data Engineer II • Tiruven Inc • Irving, TX

June 2025 – Present

- Architected the end-to-end "Carbon Footprint — CEO Insights" BI/DW platform: designed the canonical data model, built the ingestion pipelines from 10+ enterprise client systems, and delivered a unified emissions analytics layer that executives use for strategic carbon management decisions.
- Designed and deployed an autonomous AI agent in Python using synthetic datasets to simulate, analyze, and predict the key drivers behind employee efficiency and task completion — achieved 87% predictive accuracy and cut manual analysis cycles by 35%.
- Implemented AI-driven data enrichment and normalization workflows that automatically clean, standardize, and deduplicate enterprise records across disparate business domains — improving downstream reporting accuracy by 30% without manual intervention.
- Defined and maintain canonical entity schemas ensuring that a 'Customer,' 'Project,' or 'User' means the same thing across Sales, Product, and Marketing data systems — eliminating ambiguity in cross-functional reporting.
- Conduct ongoing code reviews and enforce data engineering best practices across modular transformation pipelines, ensuring testability, performance, and maintainability of the platform's logic layer.
- Provide analytical and functional support across BI/DW systems — monitoring ETL data flow, resolving performance bottlenecks, and ensuring SLA-compliant data availability for executive dashboards.

Data Engineer • Wipro Ltd • Hyderabad, TS

August 2021 – July 2023

- Built and optimized end-to-end ETL pipelines using Python and PySpark specifically to process large-scale Jira ticket log data — handling ingestion, transformation, deduplication, and loading at enterprise volume; reduced pipeline processing time by 25% and significantly improved data availability for downstream reporting teams.
- Orchestrated multi-stage data workflows using Azure Data Factory, automating data movement and transformation across cloud and on-prem environments — including scheduling, trigger-based execution, and failure recovery logic — reducing manual intervention by 30%.
- Designed and deployed 15+ reusable data intake and validation frameworks within an Agile cross-regional team, standardizing how structured data was collected and validated across business units — improving team efficiency by 20% and reducing one-off pipeline requests.
- Built Python-based REST API connectors to pull data from external systems into internal pipelines — implemented authentication (OAuth, API keys), rate-limiting, exponential backoff, and structured error handling so pipelines stayed reliable even when upstream sources degraded.
- Led CI/CD deployment of application drivers via Docker across virtualized compute environments, using Ansible and Bash for infrastructure provisioning — cut environment setup time by 40% and eliminated configuration drift across teams.
- Systematically resolved 50+ high-severity production defects through root-cause analysis of .lsc stream logs, reducing critical system downtime by 35%; authored technical documentation and trained 15+ engineers on architecture tooling and Jira workflows, cutting onboarding time by 25%.

Research Assistant • Amrita Vishwa Vidyapeetham • Bangalore, KA

March – June 2021

- Built Tableau and SQL-backed dashboards to analyze payment method adoption for a university startup — visualizations directly informed operational decisions and contributed to a 25% increase in platform utilization.
- Designed SQL-automated survey pipelines for academic research data collection — boosted response rates by 20% while maintaining compliance with institutional data governance standards.
- Trained 30+ faculty and students in SQL and data analysis workflows, improving departmental data literacy by 30% and reducing dependence on manual reporting.

EDUCATION

M.S. — Business Analytics and Information Systems — University of South Florida, Tampa, FL | May 2025
GPA 3.75 / 4.0

B.Tech — Electronics and Communication Engineering — Amrita Vishwa Vidyapeetham, Bangalore, KA | May 2021

CERTIFICATIONS

Generative AI Fundamentals — Databricks

SELECT PROJECTS

AI Chatbot — End-to-End NLP Deployment *(March 2025)*

- Built production-ready NLP chatbots using Python (spaCy, Hugging Face Transformers, TensorFlow) — handled full lifecycle from data preprocessing and model training to deployment via Flask and FastAPI; automated 20+ customer service task flows with 89% intent recognition accuracy and reduced response latency by 30%.

Credit Score Predictor — ML Pipeline *(July 2024)*

- Designed and implemented an end-to-end credit scoring ML pipeline in Python using decision tree modeling — covered feature engineering, model selection, validation, and productionization; achieved 93.7% prediction accuracy on held-out test data.

Bank Marketing Predictive Model *(October 2023)*

- Analyzed historical campaign data and client demographics to train a classification model predicting term deposit subscriptions — the model's output enabled a targeted marketing approach that improved lead conversion rates by 25%.