Harish Kesava Rao

harish.kesavarao@gmail.com | linkedin.com/in/harish-k-rao | harishkesavarao.github.io | github.com/harishkesavarao

SUMMARY

Principal Data Engineer with 12+ years building large-scale data infrastructure at Atlassian, Databricks, Amazon, Salesforce, and Indeed. Specializes in **lakehouse architecture (Delta Lake, Iceberg)**, **large-scale Spark pipelines**, and the **data foundations behind LLM and AI applications** — embeddings, vector retrieval, semantic search, and prompt pipelines.

Open-source contributor to **Apache Airflow** (author of the **DatabricksPartitionSensor**) and **Delta Lake**.

Open to senior IC, Staff, Principal, and Architect roles.

EDUCATION

M.S. Management Information Systems · University of Arizona, Eller College of Management
Tucson, AZ, USA · 2011

B.Tech. Information Technology · Anna University
India

SKILLS

AI & ML Data Infrastructure: Embedding pipelines, RAG frameworks, vector storage, semantic search, LLM prompt pipelines, Databricks Vector Search, sentence-transformers, MLflow, LangChain

Lakehouse & Storage: Delta Lake, Unity Catalog, Databricks, Neo4j, dimensional modeling (Kimball)

Data Processing: Apache Spark, PySpark, Spark Structured Streaming, Spark performance tuning, Advanced SQL, dbt

Orchestration: Apache Airflow, AWS Step Functions

Cloud & Infrastructure: AWS (EMR, Glue, S3, Kinesis, Redshift, Lambda, CDK, Athena), Azure (Event Hubs, Kafka Private Link, Storage Accounts), Terraform, Kubernetes, Docker

Languages: Python, Scala, SQL

EXPERIENCE

Principal Data Engineer · Atlassian
India · Apr 2024 – Present

- Cut **AI pipeline runtime by 86%** — from 2.5 hours to 35 minutes/day — through intermediate-step caching, ticket pre-summarization, and LLM prompt optimization, reducing compute cost alongside latency
- Shipped an **AI agent enabling 2,500 Customer Support users** to query 30K–150K monthly support tickets, ~5,000 daily Gong transcripts, and bug reports in natural language — eliminating SQL dependency for day-to-day data access
- Designed and built the **RAG framework powering the customer lens UI** — a single contextual retrieval interface across previously siloed support, transcript, and bug data, replacing manual triage workflows for support and product teams
- Delivered a **semantic search pipeline over Gong transcripts** using sentence-transformers and Databricks Vector Search — the first unified retrieval interface across all customer signals, previously inaccessible in one place
- Built a **Neo4j knowledge graph** to store and query customer signal themes, powering Customer 360 theming and analytics across Atlassian's enterprise tier
- Delivered a **natural language query interface on Databricks AI/BI** (Delta Lake + Unity Catalog) enabling non-technical stakeholders to generate insights without SQL expertise
- Owned the Data Engineering technology roadmap; established engineering best practices and guided Lead Data Engineers on architecture trade-offs across the organization

Staff Software Engineer — Data Infrastructure · Databricks
Mountain View, CA, USA · May 2022 – Mar 2024

Built core ingestion infrastructure for Databricks' internal **security detection platform** (publicly presented at *FloCon 2023* by Databricks Security) — a lakehouse-based SIEM processing security telemetry across AWS and Azure to power rule-based and ML-driven threat detection.

- Designed and shipped the **event-driven and polling connector framework** that ingests security logs from heterogeneous sources — API Gateway / Lambda / Kinesis on AWS, Event Hubs and Kafka over Private Link on Azure, S3 via SNS/SQS — into Delta Lake, all Terraform-managed and configuration-driven (per-source JSON specs for credentials, state cursors, schema, and output routing)
- Built **PySpark Structured Streaming pipelines** landing source data into the detection lakehouse for downstream feature extraction, YAML-defined rule evaluation, and MITRE ATT&CK-mapped ML detections
- Reduced data availability lag for detection workloads **from ~1 week to ~4 hours typical, with a 24-hour SLA** — enabling significantly faster threat alerting across the security platform
- Acted as **tech lead** for the ingestion workstream (2 junior data infra engineers, 1 senior data infra engineer, 1 DevOps engineer); set design-review and architecture standards adopted across the broader detection platform

Senior Data Engineer · Salesforce

Seattle, WA, USA · Sep 2021 – May 2022

- Replaced legacy ETL flows with a scalable data mart architecture modeling Tableau's end-to-end license lifecycle — reducing pipeline failures and enabling finance and product teams to self-serve license analytics
- Established **reusable ingestion patterns** that reduced time-to-onboard new license models and eliminated engineering rework for finance and product data source additions

Senior Data Engineer — Prime Video Search · Amazon

Seattle, WA, USA · Oct 2020 – Sep 2021

- Sole architect and builder of a **petabyte-scale Data Lake for Prime Video Search** on native AWS (EMR, Glue, S3, Athena, Step Functions, CDK) — delivered from zero to MVP, forming the data foundation for search ranking and ML experimentation
- Sustained consistent query performance on a **128-node Redshift cluster** under heavy concurrent analytics load at Prime Video scale through continuous optimization
- Built short-term ad-hoc pipelines and trained BI engineers and data scientists on the platform, **reducing ML experiment turnaround** and unblocking the search ranking team from core DE dependencies

Senior Data Engineer · Indeed.com

Austin, TX, USA · Oct 2017 – Oct 2020

- Eliminated cross-team metric disagreements by delivering **Kimball-modelled data marts** and automated ETL frameworks in Python and PySpark, creating a single source of truth for marketing, finance, and product analytics
- Expanded data accessibility for downstream analytics teams by architecting a large-scale PySpark/Presto/Airflow pipeline and building a **Kerberos-authenticated Presto connector** integrating previously unreachable data sources into the standard ETL framework

Senior ETL Engineer / Technical Consultant · Informatica

Austin, TX, USA · Jan 2012 – Oct 2017

- Five years across professional services and presales — distributed Informatica ETL tuning, dimensional modeling for enterprise customers, and proof-of-concept deployments supporting product evaluations and enterprise sales outcomes

OPEN SOURCE CONTRIBUTIONS

- **Apache Airflow** — Authored the **DatabricksPartitionSensor** in the official Databricks provider — merged and in production (apache/airflow); other contributions to Apache Airflow
- **Delta Lake** — Improved error observability in *SnapshotManager.getLogSegmentForVersion* — enhanced diagnostics for version resolution failures (delta-io/delta PR #5329)