

HARIKRASAD BATHINI SANKARAN

📍 Cincinnati, Ohio 📞 +1 (513) 227-1668 📩 hariprasad.sankaran@gmail.com 🌐 hariprasadbs 🌐 Hariprasad-b-s

DATA ENGINEER

SUMMARY

Result driven Data Engineer with 4 years of experience designing and optimizing large scale data architectures, pipelines, and analytical frameworks across **Databricks**, **Snowflake**, **dbt**, and **Apache Airflow**. Skilled in developing high-performance ETL/ELT workflows, implementing medallion data models, and automating data ingestion using **Python**, **PySpark**, and **SQL** across **AWS** and **Azure** environments. Proficient in building streaming and batch data solutions, optimizing pipeline scalability, cost efficiency, and reliability. Experienced in integrating security telemetry and log analytics into data platforms to support compliance and monitoring use cases. Adept at transforming raw data into actionable insights while ensuring performance, governance, and security across enterprise scale systems.

SKILLS

Data Engineering: Databricks, Python, Spark, PySpark, SQL, JSON, Database, Data warehouse, Pandas, Structured Streaming, ETL and ELT, Data Modelling, Autoloader, Medallion/Multi Hop Architecture, Unity Catalog, Cribl, Agile, Spark SQL, Data governance, Data Quality, Databricks Workflows, Apache Airflow, DBT cloud, Apache Kafka (Basics)

Cloud and DevOps: AWS (Lambda, S3), Azure (SQL Server, Azure Data Lake Storage(ADLS Gen2)), Git, GitHub, Terraform for Databricks

Programming: Python, C++, Java, REST API data ingestion, Data Structures and Algorithms in Python

Tools: Git, Github Copilot, VS Code, Cursor, AI Agents, Genie AI, Unix Commands

OS: Windows, Linux, MacOS

WORK EXPERIENCE

Data Engineer 2 - Cyber Security | Comcast Corporation | Chennai, India

Aug 2023 – Aug 2025

- Architected and maintained large scale data pipelines in **Databricks** to process multi terabyte cybersecurity telemetry from firewalls, endpoints, and threat detection systems, enabling advanced anomaly detection and security event correlation across global networks.
- Delivered end to end automated data workflows by orchestrating **Databricks Workflows** with **Terraform**, provisioning compute resources dynamically and deploying pipelines into **Unity Catalog**-managed Delta tables, improving performance and reducing data processing latency by over **30%**.
- Engineered a robust medallion (Bronze–Silver–Gold) data architecture to ensure clean, structured, and analytics ready security data for downstream analytics teams, improving query efficiency and enabling faster threat investigation turnaround.
- Implemented real time structured streaming pipelines using **Databricks Structured Streaming**, **Spark Streaming**, and **Kafka** for ingesting and processing cybersecurity logs, empowering threat hunting teams to detect suspicious activity within minutes.
- Automated multi source data ingestion using **Cribl**, **Apache Airflow**, and **AWS Lambda**, seamlessly landing security event data as **JSON** objects in **S3**, standardizing the ingestion of logs from 50+ security tools into a unified analytics layer.
- Designed a reusable framework and generic pipeline templates that reduced development effort for new data onboarding by **40%**, enabling teams to deliver new analytics feeds in days instead of weeks.
- Optimized Databricks cloud infrastructure and storage costs by \$20,000 annually through automated **Delta Vacuum**, partition pruning, and ZORDER compaction scripts for table maintenance and performance tuning.
- Developed a Proof of Concept (POC) leveraging **SparkQueryListener** and **Kafka** to monitor streaming job health, collect real-time metrics, and visualize job execution trends for proactive operational monitoring.
- Implemented Terraform modules for Databricks **CI/CD**, automating deployment of jobs, clusters, and permissions across dev, staging, and production environments, achieving consistent, reproducible releases with zero manual intervention.
- Collaborated with **cybersecurity analysts** and data scientists to design log schemas, enrich threat intelligence feeds, and operationalize anomaly detection models using **Databricks MLflow** for improved incident response accuracy.

Data Engineer | Infosys Ltd. (Microsoft Corporation Ltd.) | Chennai, India

Jun 2021 – Aug 2023

- Engineered and maintained large-scale ETL pipelines in **Azure Databricks** and **Azure Data Factory (ADF)** to ingest, transform, and aggregate multi-source datasets into **Azure Data Lake Storage Gen2 (ADLS Gen2)**, improving data consistency and refresh accuracy by over **40%**.
- Leveraged **Apache PySpark** to design high-performance distributed data processing jobs, optimizing partitioning, caching, and shuffling strategies to accelerate large-scale transformations across billions of records.
- Converted legacy **SQL/T-SQL stored procedures** into modular **PySpark ETL frameworks**, improving performance by **15%** and reducing batch job runtime by more than **30 minutes per daily load**.
- Developed a **metadata-driven ingestion framework** in **Azure Data Factory**, automating source-to-target mapping, schema validation, and dynamic pipeline orchestration—significantly increasing scalability and maintainability of production workflows.
- Implemented data lifecycle management for **Delta tables** using Python's **multiprocessing** to parallelize vacuum and optimize operations, reducing ETL maintenance windows by **30 minutes per run**.
- Designed dimensional data models (**Facts and Dimensions**) supporting self-service analytics, reporting, and executive dashboards, empowering stakeholders to drive data-backed business decisions.
- Introduced rigorous **data quality and validation checks** across all ingestion layers, ensuring integrity and consistency of business-critical data feeding BI and reporting systems.
- Developed a reusable **REST API ingestion notebook** in Python to automate external data onboarding, enabling organization-wide adoption and cutting new data source integration time by nearly **50%**.

- Collaborated with data analysts and architects to optimize query performance, standardize transformation logic, and build repeatable CI/CD processes using **Git** and **Azure DevOps**.

Data Trainee | Zoho Corporation Ltd. | *Chennai, India*

Jan 2020 – Dec 2020

- Assisted in building and maintaining **SQL-driven data pipelines** for internal reporting and analytics platforms, supporting the automation of key business metrics and dashboards used by product and marketing teams.
- Authored and optimized complex **SQL queries** to generate, validate, and troubleshoot business intelligence reports, improving report accuracy and reducing turnaround time for data requests.
- Developed **Python scripts** for data cleaning, preprocessing, and file automation, streamlining report generation and minimizing manual intervention in recurring data workflows.
- Collaborated with cross-functional teams to understand data requirements and translate them into structured data extraction and transformation processes within Zoho's internal systems.
- Gained foundational exposure to **data modeling**, **version control**, and **ETL best practices**, setting the groundwork for large-scale data engineering work in later roles at Infosys and Comcast.

PROJECTS

Real-Time Retail Demand Forecasting Pipeline (AWS) | *Personal Project / GitHub*

Ongoing

- Designed and deployed a **real-time data pipeline** using **AWS S3, Glue, and Lambda** to ingest and preprocess retail sales and inventory data, processing over **50K events per hour**.
- Built modular **ETL workflows** in **Databricks (PySpark)** to cleanse and aggregate datasets for feature engineering, improving data consistency and reducing latency by **40%**.
- Implemented a **demand forecasting model** in **AWS SageMaker** using **XGBoost**, integrating prediction endpoints with Databricks notebooks for continuous retraining.
- Automated **CI/CD deployment** with **Terraform and GitHub Actions**, provisioning AWS infrastructure and Databricks workflows across dev, stage, and prod environments.
- Delivered **Power BI dashboards** connected to Redshift for real-time demand insights, improving inventory optimization accuracy by **25%**.

Security Log Analytics and Threat Detection Platform (Azure) | *Personal Project / GitHub*

Ongoing

- Developed a scalable **data ingestion and analytics platform** on **Azure Data Factory (ADF)** and **Databricks** to process and analyze large volumes of security telemetry from firewalls, endpoints, and network logs stored in **ADLS Gen2**.
- Implemented **multi-hop medallion architecture (Bronze–Silver–Gold)** in Databricks to clean, enrich, and transform raw logs into curated datasets for SIEM.
- Automated **CI/CD workflows** using **Terraform and Azure DevOps Pipelines**, provisioning ADF pipelines, Databricks clusters with version-controlled infrastructure.
- Scheduled and monitored data ingestion and retraining pipelines using **Apache Airflow**, ensuring SLA compliance and full auditability across production runs.
- Delivered interactive dashboards in **Power BI** and **Synapse** for cybersecurity teams, visualizing real-time threat anomalies, top attack sources, and model detection metrics.

Databricks Spark Query Listener – Streaming Job Monitoring Framework | *Personal Project / GitHub*

Ongoing

- Developed a **Spark Query Listener-based framework** to capture job metrics, performance stats, and error events across structured streaming and batch pipelines in Databricks.
- Integrated **Kafka** to stream Spark listener events into a centralized monitoring layer, enabling real-time analysis and alerting.
- Implemented a **PySpark streaming pipeline** to persist monitoring data in Delta tables managed by Unity Catalog for time-series performance tracking.
- Deployed infrastructure using **Terraform** and automated **CI/CD with GitHub Actions**, ensuring consistent provisioning across environments.

EDUCATION

University of Cincinnati

Expected Graduation: Dec 2026

Master of Science in Information Technology

Cincinnati, Ohio, USA

Anna University

Jun 2016 - Nov 2020

Bachelor of Engineering in Electronics and Communication Engineering

TamilNadu, India

CERTIFICATION

- **Databricks Certified Data Engineer Professional**, Databricks
- **Databricks LakeHouse Fundamentals**, Databricks
- **Azure Data Fundamentals (DP-900)**, Azure Cloud
- **Azure Fundamentals (AZ-900)**, Azure Cloud

Issued: Nov 2025

Issued: Feb 2024

Issued: Mar 2022

Issued: Jul 2021