HARIPRASAD BATHINI SANKARAN

Cincinnati, OH | (513) 227-1668 | hariprasad.sankaran@gmail.com | LinkedIn: hariprasadbs | Portfolio

DATA ENGINEER

SUMMARY

Databricks Certified Data Engineer Professional with **5+ years** owning the full lifecycle of enterprise data platforms end to end doing architecture, ingestion, streaming, layered ELT modeling, governance, IaC, CI/CD, and cost optimization across **Databricks**, **Snowflake**, **Azure**, and **AWS**. Track record of processing **multi-terabyte daily telemetry**, unifying ingestion from **50+ sources**, cutting processing latency **30%+**, and shipping reusable frameworks adopted organization-wide that shrink new-feed delivery from weeks to days. Deep **PySpark**, **Structured Streaming**, and **Kafka** expertise complemented by **dbt** and **Apache Airflow** for warehouse-native ELT, with domain depth in **cybersecurity** (SIEM, Splunk, CrowdStrike, Cribl), **healthcare (FHIR/HIPAA)**, and **GenAI/LLM** data systems, and reporting delivered through **Power BI (Microsoft Power Platform)**.

SKILLS

Data Engineering: Python, SQL, PySpark, Apache Spark, Spark SQL, Structured Streaming, ETL/ELT, Data Modeling (Dimensional / Star & Snowflake Schema, Kimball), Data Warehousing, Data Lake / Lakehouse, Medallion Architecture, Data Quality & Governance
Warehousing & ELT: Snowflake (virtual warehouses, micro-partition pruning, clustering keys, Time Travel, Streams & Tasks, RBAC), dbt (staging → intermediate → marts layering, tests, docs), Apache Airflow (DAGs, dependencies, retries, SLAs)
Databricks Platform: Delta Lake, Delta Live Tables, Unity Catalog, Databricks Lakeflow Jobs, Autoloader, ZORDER, OPTIMIZE, Vacuum
Streaming & Ingestion: Apache Kafka, Structured Streaming, Snowpipe, Snowflake Streams & Tasks, Cribl, REST API Ingestion (retry logic, pagination), Autoloader, AWS SQS
Cloud & DevOps: Azure (Data Factory, ADLS Gen2, Azure SQL/SQL Server, Synapse, Logic Apps, Key Vault, Entra ID), AWS (S3, Lambda, Glue, SQS, Redshift, Step Functions, Secrets Manager, KMS, IAM), Terraform (IaC), CI/CD (Azure DevOps, GitHub Actions), Git/GitHub
Security & Compliance: SIEM (Splunk), CrowdStrike EDR telemetry, Cribl log routing, SOC enablement, Threat & Vulnerability Assessment, RBAC / least-privilege, column-level security, dynamic data masking, PHI/HIPAA handling
Data Formats & Stores: Delta Lake, Parquet, JSON, CSV, PostgreSQL, DuckDB, ChromaDB
GenAI & LLMs: RAG pipelines, LangChain, Ollama, vector embeddings (nomic-embed-text), prompt engineering, AI-assisted development (Claude Code, GitHub Copilot, Cursor, Gemini)
Business Intelligence: Power BI (Microsoft Power Platform), Tableau, Databricks Dashboards, Operational Analytics, Data Visualization
Programming: Python, SQL, C++, Java, Unix/Bash, SDLC coding standards, Multiprocessing/Threading, Data Structures & Algorithms

CERTIFICATION

- **Databricks Certified Data Engineer Professional (Link)** *Dec 2025*
- **Azure Data Fundamentals (DP-900) (Link)** *Mar 2022*
- **Azure Fundamentals (AZ-900) (Link)** *Jul 2021*

WORK EXPERIENCE

Student Associate - Data Analyst | ***University of Cincinnati*** **Jan 2026 – Present**

- Provided supervisor with one reliable source of truth by building and maintaining **ETL pipelines** to a single reporting database.
- Delivered **Power BI (Microsoft Power Platform)** dashboards that turn raw package center metrics into at-a-glance operational views for day-to-day team decisions.
- Reduced report generation time by optimizing **SQL** queries and data models, giving stakeholders faster access to the metrics they act on.
- Reduced manual data-verification effort with automated **Python** validation scripts that catch integrity and consistency issues.

Data Engineer 2 - Cyber Security | ***Comcast Corporation*** **Aug 2023 – Aug 2025**

- Architected and maintained large-scale **Databricks** pipelines processing **multi-terabyte daily cybersecurity telemetry** from firewalls, endpoints, and threat-detection systems, enabling advanced anomaly detection and security-event correlation across global networks.
- Reduced data processing latency **30%+** by orchestrating **Databricks Workflows** and **Delta Live Tables** with Terraform-provisioned, auto-scaling compute landing in **Unity Catalog** managed Delta tables.
- Unified logs from **50+ security tools** into a single **Snowflake** analytics layer, automating ingestion with **Cribl, Airflow, and AWS Lambda** (JSON → S3 → Snowpipe) and ending fragmented, tool-by-tool data access for analysts.
- Sped threat investigations for downstream analysts by engineering a layered **staging** → **intermediate** → **marts** model in **dbt** on **Snowflake**, delivering clean, tested, analytics-ready security data.
- Designed a log-routing architecture with **Cribl** ingesting high-volume telemetry from **Splunk**, plus custom **REST API** integrations pulling endpoint telemetry directly from **CrowdStrike**.
- Reduced SIEM noise and accelerated detection for SOC analysts by building automated **AWS SQS** routing that pushes enriched, high-fidelity threat data back into **Splunk**.
- Enabled threat-hunting teams to spot suspicious activity within minutes through **Structured Streaming** pipelines for cybersecurity logs.
- Hardened business-critical security feeds with **dbt tests** (not_null, unique, relationships) and source-freshness checks across every warehouse layer.

- Ensured data integrity across **10+ ETL pipelines** running up to **5 concurrent executions** with zero dependency failures by automating orchestration with Databricks Lakeflow Jobs and Delta Live Tables.
- Cut warehouse storage costs **$5,000/year** and compute costs **20%** through **Snowflake** warehouse right-sizing, auto-suspend/resume, clustering keys, and result-cache reuse.
- Achieved consistent, reproducible releases with zero manual intervention by implementing **Terraform** modules for Databricks **CI/CD** — deploying jobs, clusters, and permissions across dev, staging, and production.
- Accelerated ETL development cycles **30%** by integrating AI-assisted tools **(GitHub Copilot, Gemini, Claude Code)** into notebook development, validation, and documentation.

Data Engineer | ***Infosys Ltd. (Client: Microsoft Corporation)*** **Jun 2021 – Aug 2023**

- Reduced new data-source integration time **50%** organization-wide by building a reusable **Python REST API** ingestion notebook (retry logic, pagination) that other teams adopted for external data onboarding.
- Improved data consistency and refresh accuracy **40%+** by engineering large-scale ETL in **Azure Databricks** and **ADF** that ingests and aggregates multi-source data into **ADLS Gen2**.
- Cut daily batch runtime by **30+ minutes** and improved performance **15%** by re-platforming legacy T-SQL stored procedures into modular, tested layered SQL (**dbt**-pattern staging → marts) and **PySpark** frameworks.
- Accelerated transformations across **billions of records** by tuning **PySpark** partitioning, caching, and shuffle strategies for high-performance distributed jobs.
- Cut new data-source onboarding effort **40%** by building reusable pipeline templates and modular, parameterized PySpark notebooks the whole team adopted, delivering new analytics feeds in days instead of weeks.
- Increased pipeline scalability and maintainability with a **metadata-driven ADF** ingestion framework automating source-to-target mapping, schema validation, and dynamic orchestration.
- Enabled self-service analytics and executive dashboards by designing dimensional (fact/dimension) **star and snowflake schema** marts consumed in **Power BI** for reliable, decision-ready reporting.
- Hardened business-critical feeds with **data quality and validation checks** (including schema-drift detection) across all ingestion layers.

Software Engineer - Data | ***Zoho Corporation Pvt. Ltd.*** **Jan 2020 – Dec 2020**

- Improved report accuracy and cut data-request turnaround by authoring and optimizing complex **SQL** for BI reports used by product teams.
- Reduced manual effort in recurring workflows with **Python** scripts for data cleaning, preprocessing, and file automation.
- Supported automation of key business metrics and dashboards by building and maintaining SQL-driven pipelines for internal reporting.

PROJECTS

Hybrid Healthcare Data Engineering Pipeline (FHIR → Medallion → RBAC) (Ongoing) | *Personal Project* **Link**

- Engineered a Bronze → Silver → Gold pipeline ingesting synthetic **FHIR R4** bundles with **PySpark**, building dedicated Silver tables for Claims, Patients, Encounters, and Conditions.
- Combined batch (PySpark) and real-time streaming (**Python Kafka Producer → Kafka on Docker → Structured Streaming**) to simulate concurrent hospital feeds across **50 partitioned data folders**.
- Modeled the Gold layer as Kimball-compliant **dbt star-schema marts** (**2 fact, 9 dimension tables**) with deterministic **MD5 surrogate keys** and **dbt tests** (not_null, unique, relationships) enforcing referential integrity across complex FHIR reference patterns.
- Implemented **HIPAA**-aligned governance with PHI hashing and generalization — **RBAC**, column-level security, and dynamic masking enforcing field-level PHI restrictions across Admin and Clinical Analyst roles in a **PostgreSQL** serving layer, queryable via **DuckDB**.

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

- Built an end-to-end security-telemetry analytics platform on **ADF** and **Databricks**, processing firewall, endpoint, and network logs from ADLS Gen2 into SIEM-ready datasets via a Bronze → Silver → Gold medallion architecture.
- Automated infrastructure and scheduling with **Terraform, Azure DevOps, and Apache Airflow** for version-controlled, SLA-compliant, fully auditable production runs.
- Developed **PySpark** transformations to clean, enrich, and structure raw JSON security events into Parquet/Delta datasets, with schema validation and data-quality checks that flag failed loads pre-consumption.
- Delivered **Power BI (Microsoft Power Platform)** and Synapse dashboards surfacing real-time threat anomalies, top attack sources, and detection metrics for cybersecurity teams.

Local RAG Pipeline: Healthcare Patient Data Q&A with LangChain and Ollama (Ongoing) | *Personal Project*

- Built an end-to-end **RAG** pipeline using **LangChain** and locally hosted **Gemma 4 (12B)** via **Ollama**, enabling fully offline, privacy-preserving Q&A over structured healthcare data without external API dependencies.
- Engineered a structured-to-text transformation layer converting **Parquet** records (output of a PySpark ETL pipeline) into semantically rich text summaries, making structured patient data searchable via natural language.
- Built a vector ingestion pipeline using **nomic-embed-text** embeddings and **ChromaDB** as the local vector store, supporting batch ingestion of multi-part Parquet datasets with persistent storage.
- Implemented an interactive conversational interface that retrieves semantically relevant patient records from ChromaDB and passes them as grounded context to Gemma 4, reducing hallucinations and anchoring responses to real data.

EDUCATION

University of Cincinnati
Master of Science in Information Technology