KIRAN KUMAR YERROLLA

Dallas, TX | +1 (248) 763-9975 | kirankyerrolla@gmail.com | LinkedIn | Portfolio

PROFESSIONAL SUMMARY

- Data Engineer with 6+ years of experience building large-scale data pipelines, lakehouse platforms, and warehouse solutions across finance, healthcare, and enterprise teams.
- Skilled in using PySpark and Databricks to process high-volume datasets, managing Delta Lake layers on AWS S3, and developing batch and streaming ingestion with Glue, Kinesis, Lambda, Kafka, and Flink.
- Experienced in designing dimensional models, creating dbt transformations, and preparing curated datasets in Snowflake and Redshift for reporting and operational use.
- Strong background in Airflow orchestration, SQL, Python, and data quality practices, with a track record of improving pipeline speed, lowering storage costs, and enabling near real-time analytics for business stakeholders.

TECHNICAL SKILLS

Cloud Data Services: AWS (S3, Glue, Lambda, AWS DMS, Kinesis, Kinesis Firehose, EMR, IAM, Step Functions, Redshift), Azure (Data Factory, Synapse, ADLS).

Big Data & Processing: Apache Spark (PySpark, SparkSQL), Apache Flink, Databricks, Kafka, Hive, Delta Lake.

ETL / ELT & Orchestration: Airflow, AWS Glue ETL, dbt, Prefect, Azure Data Factory, SSIS, PL/SQL.

Databases & Warehousing: Snowflake, Amazon Redshift, Azure Synapse, PostgreSQL, SQL Server, Data Modeling (Star Schema, Snowflake Schema, SCD Types).

Programming & Scripting: Python, SQL, Shell Scripting.

Analytics & BI: Power BI, Tableau, AWS QuickSight.

Data Quality & Validation: Great Expectations, AWS DataBrew.

CI/CD & DevOps: GitHub Actions, AWS CodeBuild, GitLab, CircleCI.

Data Management: Batch & Streaming Pipelines, Data Quality & Validation, Metadata & Lineage, Partitioning & Optimization (Z-Order, Clustering).

AI / LLM Data Workflows: LangChain, Vector Databases (FAISS, Pinecone), Embeddings-based Retrieval, RAG Utilities, Semantic Search, AI-driven Metadata Extraction.

PROFESSIONAL EXPERIENCE

AI & Data Engineer **Nov 2023 - Present**

Rocket Mortgage | Detroit, MI

- Built automated ingestion pipelines on AWS Glue, Lambda, and Step Functions to pull structured and semi-structured data from loan servicing systems, cutting manual refresh effort and improving daily lakehouse updates by over 90 %.
- Designed Kafka- and Flink-based streaming flows to capture mortgage lifecycle events in near real time, allowing finance and servicing teams to monitor customer activity sooner and improving reporting freshness by 80 %.
- Implemented Delta Lake standards on Databricks with ACID tables, partitioned layouts, and versioned histories, ensuring reliable curated layers used for regulatory reporting and monthly operational performance metrics.
- Developed PySpark transformation routines with schema validation, SCD-Type-2 tracking, and multi-domain consolidation, reducing onboarding time for new data requirements and lowering repetitive engineering rework.
- Introduced AI-driven metadata extraction using Python embeddings and vector search to classify incoming files and identify document attributes, helping reduce manual tagging effort for ingestion teams during large-volume cycles.
- Created retrieval workflows using LangChain and vector databases to surface internal knowledge references and automate data documentation steps, which shortened analyst research cycles during reconciliation and quality reviews.
- Enhanced Snowflake and Redshift workloads by reviewing query behavior, refining clustering and distribution settings, and reorganizing model structures, improving dashboard refresh speeds and BI query runtimes across finance reporting.
- Strengthened data quality and deployment practices through Great Expectations checks, DataBrew profiling, and CI/CD pipelines in GitHub Actions and CodeBuild, lowering schema drift, stale partition issues, and release inconsistencies for downstream analytics.

Data Engineer **Sep 2022 - Nov 2023**

CVS Health | Detroit, MI

- Built PySpark pipelines in Databricks to organize pharmacy claims, eligibility, and utilization data from multiple file formats, which improved daily data quality and allowed reporting teams to rely less on manual corrections.
- Set up streaming ingestion with Kinesis feeding Databricks structured streaming so prescription and transaction records arrived within 10 minutes, giving operations more timely visibility into pharmacy activity.
- Implemented Delta Lake tables with consistent schemas, checkpoints, and table versioning to maintain stable datasets, reducing errors in downstream BI models and lowering the number of data fixes needed during refresh cycles.
- Improved Spark job efficiency by applying broadcast joins, cache strategies, and filter pruning, resulting in 20-35% faster processing on large HL7 and pharmacy datasets during peak workloads.
- Moved legacy ETL logic to AWS S3 and Databricks notebooks to retire older on-prem scripts, which reduced recurring failures and gave the team clearer logs and traceability for daily processing jobs.
- Streamlined Snowflake and Redshift ELT stages by adjusting transformation order and minimizing heavy scans, enabling dashboards to refresh more quickly and reducing compute usage across repeated reporting flows.
- Configured Airflow and Azure Data Factory workflows to handle dependencies, retries, and notifications, helping the team keep daily pipelines on schedule and reducing delays that previously impacted analytics teams.
- Worked with pharmacy analytics groups to shape curated data marts and dimensional structures that aligned to clinical and operational definitions, improving KPI accuracy and supporting more consistent reporting across business units.

Software Development Engineer **Jan 2022 - Sep 2022**

Amazon | Seattle, WA

- Developed PySpark and Hive pipelines on EMR to structure multi-terabyte operational data with partitioning, schema checks, and reconciliation steps, ensuring batches remained consistent with the processing patterns used across broader lakehouse and ETL workflows.

- Built streaming ingestion paths using Kinesis, Firehose, Lambda, and Glue to deliver operational events into S3 and Redshift staging layers within minutes, reducing dependency on nightly batch jobs by about 60 % and supporting more current analytics for internal teams.
- Configured Glue Data Catalog crawlers and schema alignment rules so raw and refined layers stayed query-ready across engineering teams, lowering schema-related interruptions during daily processing cycles by around 30 %.
- Improved Redshift performance by refining sort keys, distribution styles, and compression settings after reviewing workload patterns, which reduced dashboard query runtimes by 25–40 % for monitoring and operations teams.
- Created PySpark transformation logic with validation checks, incremental processing, and anomaly detection to maintain accuracy across high-volume workflows, reducing follow-up corrections and keeping daily data refreshes reliable.
- Introduced QuickSight dashboards backed by CloudWatch metrics to surface ingestion lag, EMR job trends, and throughput behavior, helping engineering teams identify and resolve performance issues faster.
- Tuned EMR and Spark execution parameters—including parallelism, memory allocation, and shuffle behavior—by reviewing execution logs and slow stages, achieving 15–25 % faster completion times across recurring ETL pipelines.
- Maintained CI/CD workflows in GitHub Actions and CodeBuild to promote PySpark and ingestion updates across environments, keeping deployments consistent with upstream lakehouse, warehouse, and streaming models used throughout your data platform experience.

Associate Data Engineer **Jan 2019 - Dec 2020**
Cholamandalam Investment & Finance | Hyderabad, India

- Constructed ETL routines using SSIS, PL/SQL, and Python to bring together loan, billing, and customer datasets from branch and partner systems, improving the consistency of daily finance and risk reporting outputs.
- Developed stored procedures, views, and multi-table joins in SQL Server to support billing and reconciliation cycles, reducing execution time for month-end processing by about 20% and improving data accuracy for finance teams.
- Inspected long-running SQL workloads and introduced indexing and partitioning changes to stabilize nightly refreshes, which reduced delays and ensured updated financial datasets were available on schedule.
- Transferred selected SQL Server tables to AWS with DMS and Glue jobs so refreshed datasets could be generated earlier each day, lowering dependency on on-prem infrastructure and improving availability for morning reporting.
- Redesigned SSIS transformation logic using Python and AWS Glue to simplify processing flow and improve error handling, which helped shorten ETL runtimes by around 30% across key reporting pipelines.
- Applied PL/SQL quality checks for completeness, duplication, and threshold issues across loan and payment records, reducing the number of corrections needed and improving the reliability of risk dashboards.
- Established automated ingestion steps in Python and Lambda so incoming branch files and portfolio extracts could be standardized quickly, reducing manual touchpoints and keeping daily refresh cycles predictable.
- Engaged with finance and credit teams to shape curated datasets aligned to their reporting needs, supporting more consistent views of delinquency, collections, and loan performance across business units.

PROJECTS

Cloud Data Lakehouse Platform

- Designed a lakehouse model on Databricks and S3 using Kinesis for streaming and auto-loader patterns for batch files, incorporating schema checks and table versioning to maintain consistent Delta Lake layers for analytics and operational reporting.
- Built PySpark transformations and Glue ETL flows to process CDC events with validation rules, partitioning, and metadata-driven logic, then scheduled these through Airflow to ensure predictable refresh cycles across development and production.
- Optimized Delta Lake tables and cluster settings while adding CI/CD test steps for pipeline configurations, which reduced ingestion and processing time and supported more timely dashboard updates in Redshift.

Real-Time Streaming Pipeline for Operational Analytics

- Engineered a Kafka-based ingestion layer and Flink jobs with watermarking, stateful windows, and late-arrival handling, incorporating field-level validation to ensure consistent and reliable near real-time event aggregation.
- Developed Lambda and Firehose components to enrich and route validated events into RDS and Snowflake, adding lightweight checks and table-level tests so downstream analytics received stable and ready-to-query data.
- Reduced latency from two hours to under two minutes by tuning Flink parallelism and checkpoint settings, aligning deployment steps with CI/CD workflows, and restructuring downstream tables for faster dashboard consumption.

Data Warehouse Modernization & ETL Rebuild

- Refactored dimensional models in Snowflake and Redshift by restructuring fact/dimension tables and refining SCD logic, adding data quality tests in dbt to ensure consistent structures across financial and operational datasets.
- Developed dbt models and Python transformations with reusable staging and cleansing patterns, incorporating unit tests and schema checks to keep curated layers stable for recurring reporting workloads.
- Cut Power BI refresh times by roughly 60% by tuning SQL, reorganizing semantic layers, and validating table dependencies through CI/CD steps to support smoother dashboard interactions for leadership teams.

EDUCATION

Master of Science in Computer Science **Jan 2021 - May 2022**
Texas A&M University | Kingsville, TX

Bachelor of Technology in Electronics & Communication Engineering **Aug 2015 - Jun 2019**
Jawaharlal Nehru Technological University | Hyderabad, India

CERTIFICATIONS

- Certified Data Engineer Professional - **Databricks**
- Certified Solutions Architect Associate - **AWS**
- Azure Data Engineer Associate - **Microsoft**
- Power BI Data Analyst Associate - **Microsoft**
- IBM Data Engineering Professional Certificate - **Coursera**
- Data Modeling & dbt Fundamentals - **LinkedIn Learning**