RITIK DAGAR

Senior Data Engineer | Databricks | AWS | Snowflake | LLM Pipelines | MCP | Python | PySpark
ritikdagar0203@gmail.com | +91-9958152151 | LinkedIn | GitHub | Portfolio

PROFESSIONAL SUMMARY

Assistant Vice President at Barclays with 5+ years building and shipping enterprise-scale data platforms. Own the bank's Security Posture Intelligence Platform — 500+ automated KRI/KCI metrics across phishing, vulnerability SLAs, cryptographic compliance, and MITRE ATT&CK coverage — consumed by CISO leadership and regulators weekly. Deep expertise in Databricks lakehouse architecture (Delta Lake, Unity Catalog, DLT, Structured Streaming), PySpark performance engineering, and AWS-native data systems. Actively building AI-native data infrastructure: LLM-powered semantic layers, RAG pipelines, vector-based threat detection, MCP servers for automated pipeline orchestration, and generative AI data catalogues. Recognised by the Barclays CISO organisation across 5 separate initiatives for automation excellence.

TECHNICAL SKILLS

Databricks and Delta Lake: Medallion Architecture (Bronze/Silver/Gold), Delta Lake (MERGE INTO, Z-ordering, OPTIMIZE, VACUUM, Time Travel, Schema Evolution), Auto Loader (cloudFiles, schema rescue), Unity Catalog (column-level ACLs, data lineage), Structured Streaming (watermarks, checkpointing, exactly-once), Delta Live Tables, Databricks Workflows, Databricks Asset Bundles, Terraform Provider
PySpark and Performance: Partition Strategy, Broadcast Joins, Shuffle Optimisation, Data Skew (Salting), Predicate Pushdown, AQE, Caching, Spark UI DAG profiling, RDD operations
AI and LLM Engineering: LLM Pipelines (OpenAI, Claude API), RAG (vector DB + retrieval grounding), MCP Servers (automated pipeline orchestration), Semantic Layers for BI, Vector-based Threat Detection (embeddings + similarity search), Generative AI Data Catalogues, Prompt Engineering
Cloud and Warehouse: AWS (S3, Glue, Lambda, Step Functions, Kinesis, Redshift, CloudWatch, EventBridge, IAM, Boto3, Terraform), Snowflake (Clustering Keys, Streams and Tasks, Warehouse Optimisation, Result Caching, Query Profile Analysis), dbt (incremental models, SCD Type 2 snapshots)
Data Engineering: Python (Pandas, Boto3, OOP, Modular ETL), Advanced SQL (Window Functions, CTEs, Stored Procedures, Execution Plans, Index Optimisation), SSIS, Great Expectations, Data Contracts, Schema Governance
DevOps and Observability: GitLab CI/CD, GitHub Actions, Autosys, AWS Step Functions, CloudWatch, Splunk, MITRE ATT&CK, GitHub Copilot, Claude AI

PROFESSIONAL EXPERIENCE

Barclays — *August 2021 - Present*

Assistant Vice President — Data Engineer | *Noida, India* — *2024 - Present*

- **Own the bank's Security Posture Intelligence Platform on Databricks medallion lakehouse:** Auto Loader ingests 500+ KRI/KCI metrics (phishing rates, vulnerability SLAs, cryptographic compliance, application security, MITRE ATT&CK) from APIs, SharePoint, and databases to S3 Bronze; PySpark Silver jobs apply SHA-256 deduplication and SCD Type 2 Delta MERGE; Snowflake Gold with optimised clustering serves CISO and regulatory Tableau dashboards weekly.
- **Resolved PySpark data skew (3.5h → 28min):** profiled Spark DAG via Spark UI, identified single executor bottleneck on skewed metric category key, applied salting, switched sort-merge to broadcast joins for dimension tables, enabled AQE and predicate pushdown — 87% runtime reduction, same cluster budget, zero infrastructure scale-up.
- **Building MCP server for automated pipeline orchestration:** LLM-driven orchestrator that reads pipeline metadata, detects failures, auto-generates remediation steps, and triggers Databricks Workflows — reduces on-call toil by automating first-response for 80% of pipeline alerts.
- **Built LLM semantic layer over Tableau dashboards:** LLM grounded on dashboard metadata (calculations, filters, data sources, refresh cadence) answers stakeholder questions in plain English — 'why did this metric spike?' — without requiring a data engineer in the room.
- **Enforced Unity Catalog governance across all layers:** column-level ACLs per team (CISO, Intel, Hunt, Detection), full lineage from raw source to dashboard, schema drift detection blocking breaking upstream changes before propagation — audit-ready with zero manual access management.
- **Reduced AWS platform cost by 35%:** Glue DPU rightsizing (profiled, dropped 10 DPUs to 6 where parallelism plateaus), S3 Intelligent-Tiering for cold Bronze data (Glacier after 30 days), Redshift WLM queue isolation separating dashboard from ETL load, off-peak EventBridge scheduling — broken down: 15% storage, 12% compute, 8% scheduling.
- Delivered Cyber Operations dual-mode platform for CISO, Intel, Hunt, and Detection teams — Kinesis streaming at 15-min latency for operational dashboards, nightly Glue batch for Redshift historical aggregates; Splunk extracted via Lambda every 5 mins, partitioned by threat category across 8 shards.
- People management — delivery planning, GitLab code reviews, graduate onboarding programme (fresh hires to production in 3-6 months), multiple engineers promoted within 18 months.

Senior Data Engineer | *Noida, India* — *June 2023 - 2024*

- **Led SSIS-to-Databricks migration:** established medallion architecture standards, Delta Lake table design, PySpark coding guidelines, and dbt incremental model patterns adopted across the team. Built data contract enforcement framework blocking upstream schema changes before they break downstream pipelines.
- **Structured Streaming ingestion from Kafka and REST APIs:** trigger-based micro-batch, watermark configuration for late-arriving data, exactly-once checkpointing — pipeline SLA improved from 85% to 99% via Delta Lake expectations and Great Expectations validation gates.
- Built vector-based threat detection sandbox — security telemetry embedded into vector DB, similarity search surfaces 'this resembles a prior incident' even when raw text doesn't match known signatures.

Data Engineer | *Pune, India* — *August 2021 - June 2023*

- Built KRI/KCI pipeline framework from scratch — Python and SSIS ingestion from 20+ sources, SQL Server with stored procedures and views, Autosys dependency chains, Tableau dashboards replacing manual Excel reporting. Cut pipeline execution time 35% through execution plan analysis, index tuning, and SSIS buffer optimisation.

- Introduced SHA-256 hash deduplication at Bronze + MERGE INTO Delta at Silver + Great Expectations reconciliation — duplicate incidents across 20+ source systems reduced to zero. Independently caught a live financial metric duplication bug before it surfaced in reporting.

PROJECTS — PRODUCTION AND APPLIED R&D

MCP Server for Automated Pipeline Orchestration | *Python | MCP | Databricks Workflows | Claude API | GitLab*

- LLM-powered orchestration layer that reads pipeline metadata and failure context, reasons about root cause, auto-generates and executes remediation steps via Databricks Workflows API. Removes the on-call engineer from 80% of first-response pipeline alerts. github.com/Ritik2703

Vector-Based Threat Pattern Detection | *Python | Vector DB | Embeddings | Security Telemetry*

- Security telemetry embedded into a vector database — new alerts compared against historical incidents by semantic similarity, surfacing pattern variants that rule-based detection misses. Directly extends the cyber operations domain expertise built at Barclays. github.com/Ritik2703/vector-threat-detection

RAG Assistant for Governance Documentation | *Python | RAG | Vector DB | Claude API | Policy Docs*

- Retrieval-augmented assistant over internal policy and control documentation — control owners query policy in plain English, responses grounded and source-cited. Cuts manual audit lookup time and eliminates free-form hallucination risk. github.com/Ritik2703/rag-governance-assistant

LLM Semantic Layer for BI Dashboards | *Python | LLM | Tableau Metadata API | Embeddings*

- LLM grounded on Tableau workbook metadata (calculations, filters, data sources, refresh cadence) — stakeholders ask 'what does this chart show' or 'why did this metric move' in plain English, model answers from actual dashboard logic. github.com/Ritik2703/llm-tableau-semantic-layer

Generative AI Data Catalogue | *Python | LLM | Table Metadata | Column Stats | Lineage Graph*

- GenAI layer inspecting table structure, column statistics, and upstream lineage to draft human-readable descriptions and lineage narratives — cuts documentation time to a review-and-approve step. github.com/Ritik2703/genai-data-catalog

EDUCATION

SRM Institute of Technology *2017 - 2021*
B.Tech — Computer Science and Engineering | GPA: 89.25% (First Class with Distinction)
Greenfield Public School *2017*
Class XII — PCM | 85.4%