Gaurav Kumar

Senior Data Architect | Databricks Lakehouse, Distributed Data Platforms, Streaming & Production Delivery

+91 73493 59579 | Gauravsinha0305@gmail.com | linkedin.com/in/ergauravsinha | github.com/GauravSinha-SDE | Bengaluru, India

Languages: Hindi (native), English (fluent)

SUMMARY

Senior Data Architect with 10 years delivering production data platforms on Databricks across Azure and AWS, in supply chain, manufacturing, energy and oil & gas, fuel retail, and digital advertising. Standardised a 20+ pipeline Delta Lake estate serving operations across 175+ countries, cut Spark processing time by 40%, reduced data freshness from hours to minutes, migrated SAP HANA to Snowflake with zero data loss and 30% faster queries, and built the data foundation behind a supply chain control tower that removed $30M of operating cost. Led data-platform integration planning for a $40B merger. Builds metadata-driven, configuration-based frameworks that other teams adopt as standard, governs them with Unity Catalog, and ships them as code with Terraform, Databricks Asset Bundles, and GitHub Actions. Sets scope where requirements are loosely defined and owns the result through production, incident management, and tier-3 escalation. Databricks Certified Data Engineer Professional.

EXPERIENCE

Kimberly-Clark May 2022 – Present

Senior Data Architect | *Global Data Platform Centre of Excellence* | *APAC Global Distinguished Award* *Bengaluru, India*

Domain: Consumer goods supply chain, logistics and manufacturing. **Technologies:** Databricks, Delta Lake, Unity Catalog, Lakeflow Declarative Pipelines, Databricks Workflows, Auto Loader, PySpark, Spark Structured Streaming, Terraform, Databricks Asset Bundles, GitHub Actions, Azure (Data Factory, Event Hubs, Synapse, ADLS Gen2, Key Vault, DevOps), Snowflake, SAP HANA, MLflow, Monte Carlo, pytest, Control-M, Power BI

- Own end-to-end delivery of the Azure data platform serving operations across 175+ countries, from architecture through rollout, monitoring, and tier-3 escalation; set scope directly with business and SAP functional stakeholders across APAC, EMEA, and Americas, and coach 6 engineers.
- Standardise 20+ batch and streaming pipelines on a Delta Lake medallion architecture using Lakeflow Declarative Pipelines, Databricks Workflows, ADF, and Control-M, with automated schema evolution, pre-promotion quality gates, and Monte Carlo freshness, volume, and schema monitors in production.
- Engineer event-driven ingestion on Azure Event Hubs with Spark Structured Streaming and Auto Loader, using checkpointing, exactly-once semantics, and checkpointed restartability; cut data freshness from hours to minutes for mission-critical operational reporting.
- Tune Spark through execution-plan analysis, shuffle and partition strategy, broadcast joins, file layout, and cluster sizing; ship schema changes with zero consumer downtime via atomic Delta swaps, MERGE upserts, and time-travel rollback.
- Design Unity Catalog governance across the estate: catalog and schema access models, least-privilege grants, column-level masking on sensitive attributes, and end-to-end lineage over pipeline and ML model inputs and outputs.
- Provision Databricks environments as code with Terraform and Databricks Asset Bundles, deployed through GitHub Actions and Azure DevOps CI with layered pytest suites, pre-promotion validation, and multi-environment promotion.
- Led $40B Kimberly-Clark and Kenvue merger integration planning, a SAP HANA to Snowflake migration with zero data loss and 30% faster queries, and Data Vault and dimensional modelling that replaced ad-hoc reporting with governed Power BI self-service.

Harman Feb 2021 – May 2022

Senior Data Engineer | *Clients: RaceTrac & Lebara* | *Best Employee of the Year* *Bengaluru, India*

Domain: Fuel retail and telecom. **Technologies:** Databricks, PySpark, Spark Structured Streaming, Delta Lake, AWS (MSK, Kinesis, Lambda, Glue, S3), SQL

- Engineered high-throughput ingestion and transformation pipelines on Databricks and PySpark for store telemetry, introducing Structured Streaming to cut data freshness from hours to minutes for real-time monitoring and alerting.
- Built event-driven streaming ingestion on AWS with MSK and Kinesis for high-throughput event capture, Lambda for event-triggered processing, and Glue for cataloguing and transformation over S3, with checkpointing, exactly-once semantics, and automated retry and dead-letter handling.
- Reduced end-to-end Spark processing time by 40% through execution-plan analysis, partition optimisation, broadcast-join strategy, and executor configuration at scale.
- Built a config-driven, reusable PySpark ingestion framework adopted as the standard across multiple teams, and authored SQL reconciliation logic that improved dashboard accuracy by 25%.
- Partnered with data-science teams to productionise feature-engineering pipelines on Databricks, delivering model-ready datasets that accelerated downstream model development.

Brillio Jun 2020 – Jan 2021

Data Engineer | *Client: Microsoft Bing Ads* | *Brillian of the Month* *Bengaluru, India*

Domain: Digital advertising. **Technologies:** Azure Data Factory, PySpark, Azure Data Lake, SQL

- Migrated 100+ TB of legacy data to Azure with zero data loss and a 60% reduction in ETL processing time, partnering with client stakeholders, QA, and business owners through design, cutover, and validation.
- Created parameterised ADF templates and reusable PySpark libraries adopted as the team standard, shortening delivery timelines by roughly 30%, plus SQL scripts validating source-to-target integrity at scale.

Mu Sigma Dec 2018 – Jun 2020

Data Engineer | *Industrial IoT & Predictive Maintenance* | *Client: Petrofac* | *Spot Award (x2)* *Bengaluru, India*

Domain: Energy, oil and gas; industrial IoT and supply chain. **Technologies:** Databricks, Apache Kafka, PySpark, AWS (S3, EC2, EKS, EBS, IAM, KMS, CloudWatch), Kubernetes, Docker

- Architected industrial IoT pipelines on Databricks for real-time ingestion from 20,000+ field sensors, contextualising signals (tags, units, asset hierarchies) into a unified layer for downstream AI workloads.
- Delivered the data foundation behind a supply chain control tower over asset, inventory, and maintenance signals, enabling a $30M reduction in operating cost across the client estate.
- Built Kafka streaming ingestion with production pipelines on AWS using S3 landing and archive, EC2 compute, IAM least-privilege, KMS encryption, and CloudWatch monitoring, synchronising historian and batch data with live streams.
- Integrated machine learning models into production data systems, building and operating the pipelines behind 246 deployed models with dual training and low-latency inference paths running in parallel without contention.
- Deployed containerised data pipelines on Kubernetes (Amazon EKS) with EBS-backed persistent volumes for scalable, isolated, reproducible execution.

Capgemini Nov 2016 – Nov 2018

Software Analyst | *Procurement & Sourcing Analytics* | *Quarterly Individual Star Award* *Bengaluru, India*

Domain: Procurement and sourcing analytics. **Technologies:** Azure Data Lake, SQL, Power BI

- Designed a centralised Azure Data Lake integrating ERP, SCM, and sourcing systems for procurement, finance, and IT stakeholders, with SQL data models and Power BI dashboards.
- Applied Levenshtein, Jaccard, and Cosine similarity in SQL deduplication logic, cutting supplier master-data duplicates by 40%.

KEY PROJECTS

Metadata-Driven Ingestion Framework | *Databricks, PySpark, Auto Loader, ADLS Gen2, Synapse, Event Hubs* *Kimberly-Clark*

- Built a reusable framework integrating OLTP systems, file feeds, REST APIs, and MSK, Kinesis, and Event Hubs streams into the Lakehouse with automated schema evolution and checkpointed recovery, replacing bespoke per-pipeline builds across teams.
- Drove configuration over code: new sources onboard through metadata entries rather than new pipeline code, with error handling, retry, and dead-letter paths standard across every feed.

ML Data & Deployment Pipeline on Databricks | *Delta Lake, MLflow, Unity Catalog, Azure DevOps* *Kimberly-Clark*

- Delivered the production path for ML on the lakehouse: versioned feature and training datasets on Delta with point-in-time correctness and quality gates, MLflow tracking, CI-gated promotion, and Unity Catalog lineage tracing predictions back to governed data.
- Applied least-privilege grants over model inputs and outputs and Monte Carlo monitors upstream, so feature drift and input anomalies surface before model behaviour degrades in production.

GenAI Retrieval Layer for Supply-Chain Chatbot | *Databricks, LangChain, vector DB, LangGraph* *Kimberly-Clark*

- Built the chunking, embedding, and retrieval layer behind a conversational interface to operational data, carrying source, timestamp, and domain metadata per chunk for traceable answers, with incremental refresh so only changed content is re-embedded.

SQL Pre-Flight, Snowflake Cost Optimisation | *Snowflake, Streamlit, Python, SQL* *Kimberly-Clark*

- Profiled warehouse spend to surface CTAS recreation loops, result-cache misses, and SQL anti-patterns, then shipped a Streamlit-in-Snowflake tool that flags expensive query patterns before execution, shifting cost control left to the developer.

TECHNICAL SKILLS

Languages: Python, PySpark, SQL (advanced), Scala

Databricks & Lakehouse: Delta Lake, Unity Catalog, Lakeflow Declarative Pipelines (DLT), Auto Loader, Databricks SQL, Databricks Workflows, Databricks Asset Bundles, MLflow, Databricks REST API, medallion architecture

Spark & Streaming: Apache Spark, Spark SQL, Spark Structured Streaming, execution-plan analysis, shuffle and partition tuning, join optimisation, broadcast joins, file layout, cluster sizing, checkpointing, exactly-once semantics, Apache Kafka, Amazon MSK, Kinesis, Azure Event Hubs, Hadoop, HiveQL

Governance & Security: Unity Catalog (access control, lineage, data discovery), least-privilege grants, column-level masking, Azure Key Vault, Databricks secret scopes, managed identities, service principals, IAM least-privilege

IaC, CI/CD & Orchestration: Terraform, Databricks Asset Bundles, GitHub Actions, Azure DevOps CI/CD, Docker, Kubernetes (EKS), multi-environment promotion, Databricks Workflows, Azure Data Factory, Apache Airflow, Control-M

Cloud: Azure: Data Factory, Synapse Analytics, Data Lake Storage Gen2, Event Hubs, Functions, Cosmos DB, Key Vault, Azure DevOps

Cloud: AWS: MSK, Kinesis, Lambda, Glue, S3, EC2, EKS, EBS, IAM, KMS, CloudWatch, Secrets Manager

Databases & Warehousing: Snowflake, SAP HANA, PostgreSQL, SQL Server, GCP BigQuery

Testing & Observability: pytest (unit, component, integration), Monte Carlo, Grafana, Splunk, data-quality gates, reconciliation testing, schema-drift detection

MLOps & GenAI: MLflow (experiment tracking, model registry, versioning), feature and training dataset engineering, point-in-time correctness, CI-gated model promotion, input-drift monitoring, RAG pipelines, embeddings, vector databases, LangChain, LangGraph, CrewAI

APIs & BI: RESTful API design and consumption, JSON, Parquet, Avro, Streamlit, Power BI, SSAS

CERTIFICATIONS & EDUCATION

Certifications: Databricks Certified Data Engineer Professional (October 2026), Databricks Certified Data Engineer Associate (October 2026), Microsoft Certified: Azure Data Engineer Associate, DP-203 (January 2025)

Education: B.E., Computer Science & Engineering, JSS Academy of Technical Education, 2016