SAGAR KONKANAGAON

Data Engineer

sagarkonkanagaon@gmail.com • LinkedIn • Bengaluru, Karnataka, India

PROFESSIONAL SUMMARY

Data Engineer with 5 years of experience building and optimizing large-scale, production-grade distributed data pipelines across Finance, Automotive, and Industrial IoT domains. Deep expertise in the Hadoop ecosystem (HDFS, Hive, Spark, PySpark, Spark SQL), Azure (ADF, ADLS Gen2, Databricks), and GCP (BigQuery, Cloud Composer). Proven track record of delivering end-to-end ETL/ELT pipelines, Medallion Architecture (Bronze–Silver–Gold), Data Lakehouse implementations, and Spark performance optimization at scale. Skilled in data governance, schema evolution, data quality frameworks, metadata management, lineage tracking, and audit & compliance controls. Experienced across both on-premise Hadoop environments and multi-cloud platforms, with hands-on pipeline migration, incident management via PagerDuty, and CI/CD automation using Git and Jenkins.

CORE SKILLS

Big Data & Distributed Systems: Hadoop, HDFS, Hive, Apache Spark, PySpark, Spark SQL, MapReduce, Parquet, ORC, Delta Lake

Cloud & Platforms: Azure Data Factory (ADF), Azure Databricks, ADLS Gen2, GCP BigQuery, Cloud Composer (Airflow DAGs), GCS, Databricks Lakehouse

Programming & Databases: Python, SQL, Shell Scripting, Hive QL, Spark SQL, PostgreSQL

Architecture & Data Modeling: Medallion Architecture (Bronze–Silver–Gold), Data Lakehouse, ETL/ELT Design, Dimensional Modeling, SCD Type 1 & 2, Incremental Loading

Data Governance & Quality: Data Governance, Collibra, Metadata Management, Data Lineage, Schema Evolution, DQ Rule Frameworks, Data Validation, Reconciliation, Audit & Compliance Controls

DevOps & Operations: CI/CD Pipelines, Git, Jenkins, Azure DevOps, PagerDuty (Incident Management), UC4/Automic Orchestration, Shell Scripting, Linux

PROFESSIONAL EXPERIENCE

Senior Data Engineer — Jan 2026 – Present

eBay — Bengaluru, India

- Building and maintaining PySpark and Spark SQL pipelines across eBay's Hadoop infrastructure for ingestion, transformation, and delivery to downstream analytical systems — handling high-volume datasets with data freshness and SLA adherence.
- Migrated 300+ production pipelines from legacy UC4 (Automic) orchestration to DLS (Data Lake Studio) — splitting workflows by job-plan dependencies before migration, and building new DLS pipelines from scratch on the modern orchestration layer.
- Handled the ADLC migration — building deployment artifacts and packaging/versioning JARs, replacing newer JAR builds with prior stable versions to safeguard pipeline stability during cutover.
- Executed an end-to-end 2-year historical data backfill across high-volume pipelines, ensuring completeness and consistency of reprocessed datasets for downstream analytics.
- Validated data contract sheets with the PD team — confirming schema, field-level definitions, and source-to-target mappings before onboarding pipelines to the platform.
- Built AI repositories documenting end-to-end pipeline flows — supporting automated knowledge sharing, discovery, and faster team onboarding.
- Monitoring and triaging production pipeline incidents using PagerDuty — managing on-call alerts, diagnosing Spark and Hive job failures, and driving rapid resolution to minimise downstream data impact.

Stack: *Hadoop · HDFS · Hive · Apache Spark · PySpark · Spark SQL · Python · UC4 (Automic) · DLS · Data Contracts · ADLC · PagerDuty · Git · Linux*

Senior Data Engineer (Client Engagement) — Aug 2025 – Jan 2026

RedBlackTree — Bengaluru, India

Client: Trane Technologies (HVAC / Refrigeration IoT)

Industrial IoT Telemetry Data Platform

- Designed and optimised Hadoop–Spark distributed pipelines to process high-volume HVAC telemetry data from 10,000+ connected IoT devices, enabling real-time equipment health monitoring and predictive maintenance analytics.
- Built PySpark transformation jobs on Parquet datasets in HDFS — applying window functions, aggregations, broadcast joins, and schema enforcement for structured telemetry ingestion into partitioned Hive tables.
- Implemented schema evolution handling using dynamic schema inference — adapting pipelines when IoT firmware updates introduced new or modified fields, maintaining full downstream compatibility without pipeline redeployment.
- Established data validation, lineage tracking, and metadata governance across ingestion and transformation stages — ensuring end-to-end data traceability for compliance and audit requirements.
- Collaborated with product and analytics teams to translate complex business requirements into maintainable ETL frameworks — reducing ad-hoc query load by standardising Spark-based transformation patterns across the platform.

Stack: *PySpark · Hadoop · HDFS · Hive · Spark SQL · Parquet · Python · Git · Linux*

Data Engineer **Nov 2024 – Aug 2025**

Tiger Analytics — Bengaluru, India

Domain: Automotive

Automotive Data Platform — ETL Pipeline Optimisation & Lakehouse Architecture

- Designed and optimised ETL/ELT pipelines using Apache Spark, Hive, and Azure Data Factory to process large-scale automotive manufacturing and supply chain datasets — improving end-to-end pipeline throughput significantly.
- Improved distributed pipeline performance by 35% through partition pruning, caching, broadcast joins, and Adaptive Query Execution (AQE) — reducing Spark job runtimes and cluster resource consumption.
- Built incremental ingestion workflows with watermark-based change detection using ADF triggers and Airflow DAGs on GCP Cloud Composer, eliminating redundant full-load processing and reducing data latency.
- Implemented Medallion Architecture (Bronze–Silver–Gold) on Azure Databricks with Delta Lake — building a full Data Lakehouse for progressive refinement from raw ingestion to analytics-ready datasets, with SCD Type 1 & 2 on dimension tables using surrogate keys.
- Applied CI/CD practices via Git branching and Jenkins pipelines to automate testing and deployment of Spark jobs across Dev, QA, and Production — reducing deployment time and eliminating manual configuration drift.

Stack: *Apache Spark · PySpark · Hive · ADF · Azure Databricks · Delta Lake · GCP Cloud Composer · Airflow · Python · CI/CD · Jenkins · Git*

Associate Data Engineer **Oct 2020 – Aug 2024**

XL Dynamics — Mumbai, India

Domain: Finance (Mortgage & Loan Processing)

Financial Data Processing, ETL Automation & Reporting Platform

- Developed scalable ETL workflows using Python, SQL, PySpark, and Azure Data Factory to ingest and transform mortgage, loan, and risk data from multiple transactional source systems — processing 20M+ financial records with full lineage, audit controls, and regulatory compliance.
- Implemented comprehensive data validation, reconciliation checks, and DQ rule frameworks across all pipeline stages — ensuring financial data accuracy and end-to-end traceability from source through to reporting layer.
- Wrote complex Hive queries and PySpark transformation jobs on Hadoop/Linux environments for structured and semi-structured data processing — covering cleansing, deduplication, aggregation, and SCD Type 1 & 2 handling for loan and risk dimension tables.
- Authored Shell scripts automating file ingestion, scheduling, and alerting — reducing manual operational overhead by 40%; built Tableau dashboards for loan performance and risk KPIs; produced HLD and LLD design documents for scalable solution architecture across the finance data platform.

Stack: *Python · PySpark · SQL · ADF · Hive · Hadoop · HDFS · Shell Scripting · PostgreSQL · Tableau · Git · Linux*

CERTIFICATIONS

- Databricks Fundamentals – Databricks
- Build a Data Warehouse with BigQuery – Google Cloud Skill Badge
- Build Batch Data Pipelines on Google Cloud: Cloud Composer – Google Cloud Skill Badge

EDUCATION

Bachelor of Engineering (B.E.) · National Institute of Engineering (NIE), Mysore · 2017 – 2020

OPEN-SOURCE PROJECTS

Number Passport — open-source BCBS 239 lineage and reconciliation tool **Nov 2026**

github.com/sagarkonkanagaon-netizenk/number-passport · pip install number-passport

- Built a tool that traces a reported regulatory number through SQL, dbt, PySpark and Airflow code back to its source rows, reconciling every hop to the cent with file:line evidence for each filter, join and dedup.
- Column-level lineage with sqlglot, PySpark DataFrame jobs translated step by step to SQL, and a read-only, aggregate-only privacy guard; produces a SHA-256-sealed evidence pack that can be re-verified for tampering and code drift.
- Ships as a CLI, an MCP server and an AI agent skill; CI runs the suite on Python 3.10–3.13 against real PostgreSQL and Spark, and releases publish to PyPI through trusted publishing.

Stack: *Python · sqlglot · DuckDB · PostgreSQL · PySpark · dbt · Airflow · MCP · GitHub Actions*

deja — open-source memory plugin for AI coding agents **Oct 2026**

github.com/sagarkonkanagaon-netizenk/deja

- Built a memory plugin for Claude Code and OpenAI Codex that records each failing command with the change that fixed it, and surfaces that fix the next time the error appears, in any session or repo.
- Zero dependencies and no extra LLM calls: Node's built-in SQLite with FTS5 search, memories pinned to file content hashes and flagged when stale, a data-lineage pack for SQL, PySpark, dbt and Airflow, and an offline HTML browser.

Stack: *TypeScript · Node.js · SQLite FTS5 · MCP · Claude Code hooks · GitHub Actions*