Vivek Sharma

Senior Data Engineer | Aspiring Data Architect | Engineering Leader

📍 Bengaluru, India\ 📧viveksharma9413@gmail.com\ 🔗LinkedIn

Professional Summary

A highly accomplished and forward-thinking Senior Data Engineer with 9+ years of diverse experience in architecting, designing, and deploying scalable data platforms across healthcare, fintech, SaaS, and AdTech industries. Adept at building end-to-end pipelines using cloud-native tools on GCP and AWS, orchestrating ETL workflows using Apache Airflow, and optimizing large-scale batch and streaming systems. Passionate about data modeling, automation, schema design, and pipeline observability.

Currently functioning as a Data Engineering Lead at CertifyOS, driving impactful healthcare compliance systems including NPPES ingestion, license freshness compliance, and sanctions matching. I believe in building reusable, well-documented data platforms and mentoring engineers toward excellence. Seeking to formally transition into an Engineering Manager or Data Architect role.

Core Skills

- **Cloud Platforms**: GCP (BigQuery, Composer, Dataflow, GCS, Pub/Sub), AWS (EMR, Lambda, RDS, Glue, S3)
- **Languages & Frameworks**: Python, PySpark, SQL, JavaScript, Bash, C#, Django
- **Data Engineering Tools**: Airflow, Spark, Hive, Trino, Delta Lake, Databricks, Kafka, Kinesis, Terraform
- **Modeling**: SCD Type 2, Hash-based Change Tracking, Audit Logging, Normalized/Denormalized Models
- **Streaming**: Kafka, Kinesis, GCP Pub/Sub, Spark Streaming, Firehose
- **Automation**: GitHub Actions, Jenkins, Docker, Slack APIs, Artifact Registry, CI/CD workflows
- **API & Integration**: JSON schema mapping, REST APIs, CertifyOS DAL mapping, DAG Factory
- **Visualization**: Metabase, Superset, Tableau (basic), internal dashboards
- **Leadership & Process**: Sprint planning, mentoring, architecture design, stakeholder communication

---

Professional Experience

CertifyOS

Senior Data Engineer / Acting Data Engineering Lead\ *Sep 2024 – Present | Remote (Bengaluru)*

Key Projects:

- **CB1.5 Pipeline**: Architected a fully config-driven data ingestion system capable of handling over 500 different state license formats. Reduced manual file handling to zero through dynamic schema mapping and Airflow DAG templating. Supported multi-step pipelines from GCS ingestion to BigQuery transformation to API-ready outputs.
- **SCD Type 2 Utility**: Developed a reusable BigQuery-based SCD framework that performs transactional merges using hash comparisons and timestamp tracking. Integrated across multiple sanctions datasets, ensuring auditability and data lineage.
- **NPPES to DAL Transformation**: Reverse-engineered Python record-by-record mapping logic into optimized BigQuery SQL that produces deeply nested JSON-compatible outputs for DAL API ingestion. Managed over 300+ column mappings including taxonomies, identifiers, licenses, and addresses.
- **Matching Engine**: Co-developed a fuzzy matching system for sanctions data lacking NPIs, linking records to practitioners via name, address, and license similarity. Achieved 97% attribution accuracy.
- **License Refresh Compliance**: Automated NCQA pipeline ensuring all licenses are refreshed within 30 days. Built DAGs for prioritization, scraping, enrichment, validation, and BQ ingestion.
- **GitHub Actions CI/CD**: Implemented production-ready CI pipelines for Python packages, using Google Artifact Registry. Added semantic versioning, concurrency control, retry logic, and staging/production segmentation.
- **CertifyOS API Mapping**: Designed mappings to transform tabular data into JSON payloads matching nested DAL schemas with complex field logic.

Impact:

- 🔗Provider attribution rate increased from **75% to 97%**
- ♻️ Migrated **500+ manual processors** into one dynamic DAG framework
- ⏳ Ensured **100% NCQA compliance** on license freshness
- 🔩Eliminated **manual file parsing**, enabling monthly refresh
- 📅Automated scheduler and data freshness checks with daily audit logging

Tech Stack:

BigQuery, GCS, Pub/Sub, Cloud Composer (Airflow), Python, SQL, GitHub Actions, Docker, JSON Schema, DAG Factory, REST API Integration

Leadership:

- Mentored 2 engineers, 3 analysts; onboarded cross-team collaborators
- Aligned data systems with Platform, Scraper, and API Engineering teams
- Delivered architectural designs to senior leadership including Directors & Architects

---

6sense

Senior Data Engineer – Big Data Platform\ *Oct 2022 – Sep 2024 | Remote / Bengaluru*

Key Projects:

- **Singlestore ETL Framework**: Unified batch and CDC ingestion into Singlestore with automated DAG generation, schema validation, and archival. Supported production, staging, and dev environments with cluster-safe locking.
- **Hive Table Cloning Utility**: Developed tooling to clone partitioned Hive tables into lower environments with fine-grained filter controls, schema diffs, and rollback support.
- **Data Extractor Utility**: Spark-based utility for MySQL/PostgreSQL/Singlestore extraction and S3 archival. Enabled selective partitioning, deduplication, and data retention rules.
- **Delta Batch Job Optimization**: Rewrote and tuned Spark jobs to optimize shuffle and partition strategy. Reduced lead enrichment latency from 45min → 18min.
- **Data Contract Collaboration**: Worked across Marketing Ops, Analytics, and Product to standardize schema contract versioning and SLAs, reducing breakages by 80%.

Impact:

- ⏱️Reduced latency across Spark/Delta jobs by **40%**
- ⚖️Implemented schema SLAs for 4+ cross-functional teams
- 💰Reduced S3/HDFS costs by archival tiering and lifecycle enforcement
- ✉️Streamlined table debugging turnaround from 2h → 15min

Tech Stack:

PySpark, Hive, Trino, Delta Lake, Singlestore, Kafka, Airflow, MySQL, PostgreSQL, AWS (Lambda, S3, RDS)

Leadership:

- Owned 3 internal tools adopted across 5+ platform teams
- Collaborated with Staff+ Engineers on multi-region ingestion benchmarking
- Supported dev onboarding with utilities and documentation

InCred Financial Services

Tech Lead – Data Engineering\ *Apr 2021 – Sep 2022 | Bengaluru*

Key Projects:

- **CDC Platform**: Streamed data from MySQL/Postgres/DynamoDB → S3 using AWS DMS, Kinesis, and SQS. Ingested into Databricks Delta Lake via Autoloaders for near real-time analytics.
- **Lakehouse Modeling**: Designed partitioned Delta Lake structure with retention, auditing, and versioning support. Enabled data scientists to build features using MLFlow.
- **Databricks Optimization**: Tuned Spark clusters, job retries, alerts, and pooling to reduce runtime and SLA violations.

Impact:

- 📉Infra spend down **30%** via pool tuning + job schedule optimizations

- ⏳Job latency improved **30%**, faster data refresh
- 🛡️Uptime improved with Firehose → S3 ingestion
- Mentored a 4-person data team, defined coding + review standards

Tech Stack:

Databricks, PySpark, Delta Lake, DMS, Kinesis, SQS, S3, DynamoDB, MLFlow, Presto, Metabase

---

Slice

Lead Data Engineer → Data Engineer\ *Nov 2019 – Apr 2021 | Bengaluru*

Key Projects:

- **Slice Data Lake**: Migrated analytical queries off MongoDB by architecting PySpark → S3 → Redshift system using Kinesis + Glue jobs.
- **Neptune Risk Engine**: Designed AWS Neptune-based User Graph linking users to risk/revenue traits; directly impacted credit underwriting logic.
- **Jarvis Notification System**: Slack-based alerting framework powered by SNS and S3 triggers. Plug-and-play adoption by product & analytics teams.
- **Data Propagator**: Created safe data push framework for PMs to write into RDS/Redshift.

Impact:

- 💰Saved \$10K/month infra costs + \$600/month on report tooling
- 🕐Enabled self-serve reporting across PMs, analysts
- 🛡️Prevented fraud by linking "risky" users via User Graph
- 💼Built and led a team of 3 data engineers

Tech Stack:

AWS EMR, Glue, Redshift, Neptune, SNS, Slack APIs, Python, PySpark, S3, RDS, MongoDB

Particle41 India LLP

Software Developer → Data Engineer\ *Nov 2016 – Oct 2019 | Pune*

Key Projects:

- **Onboarding Engine**: Built Spark + Hive pipeline to process client CRM data for identity onboarding.
- **Identity Graph**: Used GraphFrames on Spark to unify sessions, cookies, emails into user profiles.
- **Ad ROI Model (Vevo)**: Developed MLLib rule-based targeting model for Vevo ad delivery.
- **B2B Catalog (RepSpark)**: Built React + C# interface for wholesalers to place real-time orders.
- **Mobile UX (MiMedia)**: Enhanced Android onboarding with animated walkthrough screens.

Impact:

- 🩲Served 10+ enterprise clients with onboarding ETLs
- 🧦Praised by U.S. client CTO for leadership & reliability
- 🎓Mentored 3 new devs during team expansion
- 💡 Unified multiple systems into integrated platforms

Tech Stack:

AWS EMR, Spark, Hive, GraphFrame, ReactJS, Python, Java (Android), C#, PostgreSQL, Redshift

---

Persistent Systems (Intern)

Project Intern\ *Sep 2015 – Jul 2016 | Pune*

- Built Django-based stock prediction app using Yahoo Finance APIs + TA indicators
- Delivered project across timelines, managed client demos and documentation
- Co-authored research paper: *"Stock Market Forecasting Using Hybrid Methodology"* in IJIR 2016\ Read Publication

---

Education

Degree	Institute	Year	Grade
B.E. in Information Technology	University of Pune	2016	73.08%
12th (HSC)	Kendriya Vidyalaya No.3, Pune	2011	82.2%
10th (SSC)	Kendriya Vidyalaya No.3, Pune	2009	91.2%

Certifications

- Introduction to Big Data — Coursera (Oct 2019)
- SQL Fundamentals — SoloLearn (May 2016)

Awards & Achievements

- ACM ICPC Amritapuri Onsite Qualifier (2014)
- Global Rank #242 – TCS CodeVita 2014
- Facebook HackerCup & Google CodeJam Qualifier
- Winner – C Programming @ SAE Pune (2013)
- U19 KV Cricket Mumbai Region, Chess Runner-up (2013)

Publications

- **Stock Market Forecasting Using Hybrid Methodology** — IJIR, May 2016\ Read Paper

Leadership & Community

- HackerRank Campus Ambassador & Coder Club Founder
- Tech Fest Head Coordinator (2015–2016)
- Student Council Member, KV Pune (2014–2015