

# Vivek Sharma

Senior Data Engineer | Aspiring Data Architect | Engineering Leader

 [your.email@domain.com] |  [LinkedIn] |  [GitHub / Portfolio Site]

---



## Professional Summary

Senior Data Engineer at **CertifyOS**, leading the design and delivery of large-scale healthcare compliance and provider attribution systems. Though not yet titled, I actively function as a **Data Engineering Lead**, architecting pipelines, mentoring engineers, and aligning across cross-functional teams. With deep expertise in GCP, BigQuery, Airflow (Composer), and API integration, I've built reusable, automated systems that power critical workflows — from NPPES integrations to NCQA license compliance and sanctions ingestion. Now seeking to step formally into an **Engineering Manager** or **Data Architect** role where I can scale systems, people, and outcomes.

---



## Core Competencies

- **Data Architecture & Modeling:** SCD Type 2, temporal tracking, config-driven ingestion
  - **GCP Stack:** BigQuery, GCS, Pub/Sub, Cloud Composer, Dataflow
  - **Languages & Tools:** Python, SQL, Bash, GitHub Actions, Docker
  - **API & Schema Integration:** Nested JSON mapping, CertifyOS DAL, NPPES, license payloads
  - **ETL & Orchestration:** Airflow (Cloud Composer), Pub/Sub triggers, artifact versioning
  - **Leadership:** Sprint planning, stakeholder communication, mentoring
  - **Cross-Functional Collaboration:** Platform, Pipeline, Scraper, and Data Science teams
- 



## Team & Leadership Involvement

- Present architectural designs to **Engineering Managers, Directors, and Chief Architects**
  - Participate in **sprint planning**, story grooming, and defining deliverables with PMs
  - Lead technical decision-making across Data, Platform, and Scraper teams
  - Mentor junior engineers and analysts on pipeline design, BigQuery modeling, and automation
  - Drive best practices in CI/CD, schema evolution, and DAG observability
- 



## Ongoing & Recent Projects

Project	Role	Description
<b>CB1.5 Pipeline – Sanctions &amp; License Ingestion</b>	Lead Engineer & Architect	Designed and led CertifyOS's largest ingestion system, migrating 500+ manually processed license file types into a unified, job-config-driven pipeline using Cloud Composer and BigQuery. Achieved full automation of monthly state license refreshes, previously handled via separate scripts per file.
<b>SCD Type 2 Utility</b>	Designer & Developer	Built a reusable SQL framework for Slowly Changing Dimensions (Type 2) using hash-based change tracking. Applied across sanctions datasets lacking primary keys, ensuring accurate historical tracking and auditability with isCurrent, startDate, and endDate logic.
<b>NPPEs to DAL Transformation</b>	Architect & Engineer	Developed a fully parity-matched transformation layer in BigQuery SQL to convert flat npes_raw tables into nested DAL-ready practitioner payloads for API ingestion. Coordinated with Platform team to validate schemas, manage chunked ingestion, and support 300+ column mappings.

<b>Matching Engine for Sanctions Attribution</b>	Co-designer	Co-developed a fuzzy matching engine to attribute provider sanctions to practitioners using name, address, and license matching. Critical for linking records lacking NPIs, improving attribution rate from 75% to 97%.
<b>Scraper Automation Layer DAG</b>	Builder & Coordinator	Created an orchestration DAG to automate ingestion of scraper outputs into validation pipelines. Integrated with license enrichment and matching layers to support NCQA license freshness compliance. Worked closely with the Scraper team to finalize contracts and ensure reliable ingestion.
<b>NCQA License Refresh Pipeline</b>	Designer	Enabled compliance by designing a system that ensures all licenses are refreshed within 30 days, powered by the scraper automation layer. Tracked freshness across license types and ensured seamless updates into internal compliance systems.
<b>GitHub Actions Release Pipelines</b>	DevOps Contributor	Implemented CI/CD workflows for Python package deployment to Google Artifact Registry with retry logic, version bumping, and concurrency controls. Segregated staging (release/stage-*) and production (release/main) workflows with semantic versioning and selective rebuilds.
<b>CertifyOS API Mapping &amp; Integration</b>	Technical Lead	Mapped multiple flat datasets (NPPES, sanctions, licenses) to deeply nested API schemas. Ensured complete schema validation, null handling, type coercion, and custom field logic for payload ingestion into DAL endpoints.
<b>State Sanctions Dataflow Design (CB1.5)</b>	Dataflow Architect	Designed the DataFlow logic for ingesting state sanctions files using a config-driven approach with schema and transformation logic maintained in BQ config tables. Enabled scalable file parsing and publishing of standardized records into Pub/Sub for downstream CB2.0 systems.

---



## Key Impact Metrics

Area	Impact
Provider Attribution	↑ Improved from <b>75% to 97%</b> using enriched matching engine
License Refresh	🚀 Migrated <b>500+ script-based processors</b> to a single automated pipeline
Compliance (NCQA)	🕒 Ensured license data is refreshed <b>within 30-day SLA</b>
Manual Effort	✗ Reduced manual touchpoints to zero via config-driven pipeline design
Code Reuse	♻️ Delivered reusable frameworks adopted across multiple datasets and teams



## Education

**Bachelor of Engineering (Information Technology)**

*University of Pune, India – 2016*




## Career Goal

To transition formally into an **Engineering Manager** or **Principal Data Architect** role — combining deep technical vision with team leadership to build scalable, resilient data platforms that drive critical business outcomes.



# 6sense

## Senior Data Engineer

 Oct 2022 – Sept 2024 | Bengaluru, India (Remote/Hybrid)

---



### Summary

At 6sense, a leading AI-driven Account-Based Marketing (ABM) and Revenue Intelligence platform, I served as a Senior Data Engineer within the Big Data Platform team. I helped design and implement scalable internal tools and frameworks used across 6sense’s real-time and batch data systems — powering customer-facing insights, lead enrichment, and campaign optimization. I contributed to platform-wide initiatives supporting one of the **largest big data clusters in the industry**, with over **1 million jobs processed daily** across Spark, Trino, and Delta Lake pipelines.

---



### Key Projects

Project	Role	Description
Singlestore ETL Framework	Lead Developer	Led development of a unified ETL framework for managing ingestion, transformation, and archival of real-time and batch data into 6sense’s Singlestore warehouse. Features included auto-scaling, CDC support, Trino-Hive integration, and cross-region deployments. Enabled seamless real-time access and large-scale batch reporting.
Data Extractor Utility	End-to-End Owner	Built a generic Python + PySpark-based utility for extracting and archiving data from MySQL, Postgres, and Singlestore into S3/HDFS. Allowed teams to configure filtering,


partitioning, and retention rules without code changes — streamlining compliance and ops workflows.


<b>Hive Table Cloning Utility</b>	Lead Developer	Created a utility to rapidly clone production Hive tables to dev environments for staging/debugging. Reduced manual time from hours to minutes with built-in error handling and flexible config support.
<b>Delta &amp; Spark Batch Job Optimization</b>	Contributor	Improved performance of legacy Spark jobs handling TB-scale lead enrichment data. Reduced job run time and optimized partitioning and shuffle logic across multiple data domains.
<b>Cross-Team Data Contract Ownership</b>	Collaborator	Worked with Product, Marketing, and Analytics teams to manage data schema consistency, ingestion failures, and format evolution — reducing downstream breakages and improving SLAs.




## Key Impact Metrics

Area	Impact
<b>Platform Efficiency</b>	⌚ Reduced ingestion and job latency across Spark & Delta jobs by tuning partitioning and I/O
<b>Developer Productivity</b>	🚀 Enabled rapid table cloning and pipeline debugging across staging environments

**Data Consistency**  Standardized schema contract validation across cross-functional teams

**Cost Optimization**  Participated in AWS S3/HDFS archival strategy to reduce storage & query costs

**ETL Framework Adoption**  Internal tooling adopted org-wide for managing batch and CDC ingestion pipelines

---



## Tech Stack

- **Languages:** Python, PySpark, SQL
  - **Compute & Storage:** Singlestore, Hive, Trino/Presto, Delta Lake, Spark, HDFS, S3
  - **Databases:** MySQL, PostgreSQL, RDS
  - **Streaming & Messaging:** Apache Kafka
  - **Cloud Infrastructure:** AWS (S3, Lambda, RDS, Kinesis)
  - **Orchestration & Monitoring:** Custom internal tools, Spark schedulers
- 



## Collaboration & Responsibilities

- Served as the **Lead Developer** on key internal utilities used across the Big Data team
- Collaborated with Staff Engineers and Platform Leads on **multi-region replication, CDC strategies**, and **performance benchmarking**
- Worked cross-functionally with **Marketing Ops, SalesOps, and Data Science** to meet SLA, contract, and delivery needs
- Participated in monitoring and maintaining high-throughput Spark and Trino jobs that support 6sense's ABM intelligence layers
- Supported platform-wide initiatives to reduce AWS resource usage and improve system resilience





# InCred Financial Services Ltd

## Tech Lead – Data Engineer

📍 Apr 2021 – Sept 2022 | Bengaluru, India

---



### Summary

At InCred, a digital-first lending and financial services provider, I served as the **Tech Lead for the Data Platform team**, driving the architecture and development of a robust, real-time data ecosystem that powered analytics, credit risk modeling, and reporting across the business. I owned end-to-end data platform initiatives—from ingestion to lakehouse modeling—while mentoring engineers and analysts, optimizing infrastructure, and establishing data reliability and cost efficiency as core principles.

---



### Key Project

Project	Role	Description
InCred Data Platform	Tech Lead & Platform Architect	Designed and led the development of a near real-time data platform ingesting CDC data from <b>MySQL, Postgres, and DynamoDB</b> using <b>AWS DMS, Kinesis, and SQS</b> , and transforming it in <b>Databricks</b> using <b>PySpark and Autoloaders</b> . Delivered clean, queryable Delta Lakes in S3 to power analytics, ML pipelines, and dashboards. Led cost optimization efforts that <b>reduced infra spend by 30%</b> , mentored a team of 4, and implemented monitoring, retry, and resource-tuned Databricks pools to minimize job latency.

---



## Key Impact Metrics

Area	Impact
Infrastructure Cost	💰 Reduced streaming + ETL cost by over <b>30%</b> through optimized pipeline design
ETL Latency	⚡ Cut job start time and report refresh delays by <b>30%</b> with Databricks pool tuning
System Reliability	✅ Migrated unstable on-prem Kafka-Dynamo pipelines to <b>AWS-native Kinesis Firehose</b> , improving uptime and observability
Team Growth	🧑 Mentored and upskilled a <b>team of 4+</b> engineers/analysts across PySpark, streaming, and data contract practices



## Tech Stack

- **Languages & Processing:** Python, PySpark, Spark Streaming
- **Data Engineering Platforms:** Databricks (Jobs, Autoloader, MLFlow), AWS DMS
- **Streaming Systems:** AWS Kinesis, Kafka, SQS
- **Storage & Lakehouse:** Delta Lake on S3, EFS
- **Databases:** MySQL, PostgreSQL, DynamoDB
- **Monitoring & Scheduling:** Databricks jobs, alerting, retry logic
- **Analytics:** Presto, Metabase
- **ML Collaboration:** Supported feature stores in Druid; integrated with MLFlow model tracking pipelines



## Leadership & Collaboration

- Acted as **Tech Lead and Platform Owner** for the company-wide data platform
- Solely responsible for **architecting CDC ingestion workflows** and tuning streaming pipelines for stability and cost efficiency
- Partnered with ML teams to build **feature stores** and ensure high-frequency data availability for model training
- Enabled Risk, Analytics, and Product teams to access **reliable, fresh, and performant data**
- Mentored junior engineers in **streaming design patterns, Databricks optimization, and infrastructure monitoring**



# Slice

Data Engineering Lead → Data Engineer

📍 Nov 2019 – Apr 2021 | Bengaluru, India

---



## Summary

At Slice, a fast-scaling consumer credit and payments startup, I transitioned from an individual contributor to **Data Engineering Lead**, building the foundation of the company's data infrastructure, warehouse, and tooling. I architected core systems like the **Data Lake**, **User Graph**, and **data propagators**, while also mentoring the growing data team and owning stakeholder relationships across risk, analytics, and product teams. My contributions helped **de-risk infra scale-up**, **enable self-serve analytics**, and **cut costs** on third-party tools by driving internal platform innovation.

---



## Key Projects

Project	Role	Description
<b>Slice Data Lake</b>	Lead Developer	Architected and implemented a PySpark + S3-based data lake that offloaded analytical workloads from MongoDB to Delta-compatible Parquet storage on AWS. Enabled real-time ingestion using <b>Kinesis and Glue/EMR</b> , freeing MongoDB for high-throughput API workloads and reducing infra costs by ~\$10K/month.


<b>Slice User Graph</b>	Lead Developer	Built a graph-based analytics engine using <b>AWS Neptune</b> to identify clusters of users with shared risk/revenue traits. Powered user approval/rejection logic by analyzing links to “good” or “risky” users. Improved underwriting strategies and fraud detection accuracy.
<b>S3 to Redshift/RDS Propagator</b>	Lead Developer	Developed a user-friendly tool for analysts and PMs to safely push data to <b>Redshift and RDS</b> without needing engineering intervention. Included validation, schema mapping, and write protections to avoid data corruption.
<b>Mongo to Redshift Pipeline</b>	Lead Developer	Built a CDC ingestion pipeline for transactional and behavioral user data from MongoDB to Redshift. Enabled deeper analysis by the Risk team and eliminated dependency on expensive third-party tools (~\$600/month saved).
<b>Jarvis: Slack Notification Framework</b>	Lead Developer	Created an internal Slack-based alerting system triggered by <b>SNS, S3 events</b> , and job outcomes. Widely adopted across teams for real-time ops visibility and job health monitoring. Plug-and-play for any Slack channel.



## Key Impact Metrics

Area	Impact
<b>Infrastructure Cost</b>	💰 Saved ~\$10,000/month by offloading MongoDB and ~\$600/month from third-party tool replacement
<b>Platform Reliability</b>	✅ Improved real-time ingestion stability and freed MongoDB for API performance


### Analytics Enablement

 Empowered analysts and PMs to push/report data without engineering dependency

### Fraud Risk Mitigation

 Enabled graph-based user linking for approval/rejection logic

### Team Growth

 Built and led the data engineering team from scratch, mentoring new engineers on AWS and Spark tech stack

---



## Tech Stack

- **Languages:** Python, PySpark
  - **Data Infra:** AWS S3, Glue, EMR, Kinesis, Athena, Redshift, RDS, EC2, DynamoDB
  - **Databases:** MongoDB, AWS Neptune
  - **Eventing & Messaging:** AWS SNS, S3 Events, SQS
  - **Alerting & Observability:** Slack APIs, Custom Slack Bot Framework (Jarvis)
- 




## Leadership & Collaboration

- **Promoted** from IC to **Data Engineering Lead** within 12 months based on technical ownership and delivery
- Led a **team of 3 engineers** while working directly with analytics, product, and risk teams
- Designed team processes and mentored junior members on AWS tools, infrastructure best practices, and PySpark
- Played a critical role in **selecting and deploying foundational data systems**, balancing cost, speed, and reliability
- Drove technical blogs to document internal tooling (e.g., **Jarvis**, **Data Propagator**, **User Graph**)



# Particle41 India LLP

## Software Developer

 Nov 2016 – Oct 2019 | Pune, India

---



## Summary

At Particle41, my first role post-college, I progressed from an individual contributor to a **lead developer and systems architect** across multiple domains — from **big data processing and identity resolution** to **ML pipelines and full-stack applications**. I collaborated directly with international clients, including U.S.-based CTOs, to deliver production-grade solutions. My hands-on work across **AWS and GCP**, combined with **cross-functional ownership**, helped establish a strong technical foundation that continues to support my engineering leadership today.

---



## Key Projects

Project	Role	Description
Onboarding Engine (IDify)	Lead Developer & Architect	Designed and implemented a scalable Spark-based ETL pipeline on <b>AWS EMR</b> , processing CRM data from partner clients into digital user profiles. The system included a matching engine, tag server, and reporting layer, enabling high-throughput identity onboarding and significantly enhancing online audience match rates.

<b>Identity Resolution (ID Graph)</b>	Lead Developer & Architect	Developed a user identity graph using <b>GraphFrame</b> on Spark to connect identifiers like cookies, devices, and emails into unified user profiles. Improved downstream personalization and attribution by enhancing entity linkage.
<b>DFP ROI Enhancements (Vevo)</b>	Team Lead	Led a team to build a rule-generation ML pipeline using <b>PySpark + MLLib</b> , improving Vevo's ad targeting performance and on-target impression ROI.
<b>Program Management Report</b>	Developer	Delivered an ETL system to ingest third-party buyer reports and generate automated revenue dashboards using <b>Python (PETL)</b> and <b>Redshift</b> . Also gathered client requirements firsthand.
<b>MiMedia FTE (Android)</b>	Developer	Enhanced the user onboarding experience in the Android app using animated FTE screens, improving engagement and retention.
<b>Online Catalog Viewer (RepSpark)</b>	Lead Developer	Led the development of a React + C# web app to help wholesalers build interactive digital product catalogs and take B2B orders in real time.



## Key Impact Metrics

Area	Impact
<b>Data Product Scale</b>	⚙️ Delivered Spark- and EMR-based pipelines that scaled with onboarding growth



**Client Satisfaction** 🧡 Worked directly with the CTO of ALC (a U.S.-based client), who credited me as “*a critical part of our success*” due to my reliability, technical depth, and leadership in onboarding and incident resolution

**Team Enablement** 👥 Trained new engineers and led delivery across multiple team transitions

**System Design** 🏗️ Progressed from developer to platform architect across multiple distributed systems

**Cost Optimization** 💡 Replaced manual workflows and third-party tooling with custom-built, efficient solutions



## Tech Stack

- **Big Data & Processing:** Apache Spark, Hive, GraphFrame, PySpark
- **Cloud Platforms:** AWS EMR, S3, EC2, Glue, Lambda, Athena, GCP (limited scope)
- **Languages & Tools:** Python 2.7/3, PETL, Java (Android), C#.NET, ReactJS
- **Databases:** PostgreSQL, DynamoDB, Microsoft SQL Server, Redshift
- **ML & Analytics:** Spark MLlib
- **Dev & Deployment:** Slack APIs, CI pipelines, documentation ownership



## Responsibilities & Growth

- Transitioned from graduate hire to **lead developer and system designer** within 2.5 years
- Mentored peers and junior engineers, particularly during team transitions and onboarding
- Directly engaged with global clients (U.S. and EU), understanding requirements and delivering production-quality solutions

- Received praise from **Rick Landsman (CTO of ALC)** for playing a crucial role in the success of mission-critical systems and for consistently leading incident resolution, daily team syncs, and onboarding
- Owned technical delivery across identity resolution, onboarding engines, analytics reporting, and full-stack UI projects

✓ Here's your updated **Lifelong Learning** section, now including both your certifications, publication, and early career foundation:

---



## Lifelong Learning & Foundations



### Certifications

- **Introduction to Big Data** – Coursera | Oct 2019  
[View Certificate](#)  
*Explored big data architecture, processing frameworks like Hadoop and Spark, and distributed data pipelines.*
  - **SQL Fundamentals** – SoloLearn | May 2016  
[View Certificate](#)  
*Built core skills in SQL, data querying, joins, aggregations, and schema operations.*
- 



### Publications

- **Stock Market Forecasting Using Hybrid Methodology** – *Imperial Journal of Interdisciplinary Research (IJIR)*, May 2016  
[Read Publication](#)  
*Co-authored a research paper on forecasting stock trends using statistical and machine learning models.*
- 



### Early Career Foundation

- **Project Intern – Persistent Systems, Pune** | Sept 2015 – July 2016  
*Worked on a sponsored academic project titled “Stock Market Prediction using Technical Analysis.” Developed the system using Django, HTML, CSS, and JavaScript. Gained firsthand experience in project scoping, timelines, and real-world workplace culture.*



## Education

Degree	Institution	Year	Grade
Bachelor of Engineering (IT)	Pune University	July 2016	73.08%
Senior Secondary (12th)	Kendriya Vidyalaya No. 3, 9BRD, Pune	2010	82.20%
Secondary (10th)	Kendriya Vidyalaya No. 3, 9BRD, Pune	2008	91.20%

---







## Key Achievements & Technical Competitions

- **ACM ICPC Amritapuri Regional Onsite Contestant** – 2014
  - **1st Prize** – C Programming Contest (SE), Sinhgad Academy of Engineering, Pune (2013)
  - **2nd Prize** – CodeChef Challenge, AISSMS College Fest (2014)
  - **3rd Prize** – C-Athlon, Melange Fest, BRAC's VIT, Pune (2014)
  - **Finalist** – C-Maestros, MITCOE Tesla 2014
  - **Finalist** – Oracle Think.com Web Page Contest (2007–08)
  - **Qualified** – 1st Rounds of **Facebook HackerCup** and **Google CodeJam**, 2015
  - **Global Rank #242** – TCS CodeVita 2014
  - **India Rank #619** – HackerRank CodeSprint India 2014
  - **Waitlisted Finalist** – CSI National Programming Contest 2014
  - **Runner-Up** – Chess Tournament, INNOFEST 2013, Sinhgad Academy of Engineering
  - **U-19 Cricket Player** – Represented Kendriya Vidyalaya Sangathan, Mumbai Region (2011–12)
- 



## Leadership & Volunteer Experience

-  **HackerRank Campus Ambassador** and *Founder of Coders Club* on campus
-  **Student Council Member** – 2014–15 academic year
-  **Head Student Coordinator** – *Freaky 'C'oders*, TECHTONIC 2015
-  **Head Event Coordinator** – *Freaky 'C'oders*, TECHTONIC 2016