Shril Kumar

github.com/shril
linkedin.com/in/shril

Email: shril.iitdhn@gmail.com
Mobile: +91-8521727259

Data & AI Infrastructure Engineer

Data and AI infrastructure engineer with 8+ years building large-scale data platforms, privacy and governance systems, agentic workflows, and reliable distributed pipelines. Experienced in PII classification, provenance-aware synthetic data, lineage, policy-enforced access, data lifecycle management, and production systems operating across petabyte-scale estates.

Experience

- **Roku** — Bangalore, IN
  Senior Software Development Engineer (Data) — *September 2024 – Present*
  Tech Lead – 4 engineers reporting informally
  - **PII Discovery and Classification**: Built an LLM-powered PII discovery and classification platform for Roku's **40 PB Ads Data Platform**, replacing a planned **$150K/year** vendor product; classified schema metadata into sensitive-data categories and routed predictions to legal analysts for human validation.
    - **Evaluation-to-Data Flywheel**: Clustered analyst corrections into **6 failure modes**, generated provenance-tagged synthetic examples for weak categories, and served them through retrieval-augmented few-shot prompting—increasing precision from **84% to 91% at recall** $\geq 99\%$. Established the platform as a cornerstone of Roku's data-governance strategy and began expanding it to an additional **60 PB** across core, enterprise, SQL, and NoSQL data estates.
  - **Ads Data Platform Architecture**: Re-architected hourly Spark pipelines processing **200M ad impressions and 40M conversion events daily** under a one-hour SLA; replaced mutable tracker state with stateless, idempotent Iceberg processing using partition-held state, Bloom filters, and positional deletes.
    - **Processing Reliability and Iceberg Maintenance**: Added transactional write-audit-publish controls and safe reruns to impression and conversion processing, cutting job runtime **35%**, Trino query latency **20%**, and incident recovery from **1 day to 2 hours** without increasing ingestion latency. Built an Airflow-integrated maintenance framework for these Iceberg tables and rolled it out to **100+ production tables** for compaction, snapshot management, and orphan cleanup.
  - **Data-Query MCP**: Architected a governed data-access layer for AI agents and engineers across approximately **5,000 Hive, Iceberg, and Snowflake tables**; used DataHub metadata to enforce schema, lineage, ownership, identity context, and metric-certification policy before executing queries through Trino. Implemented per-user OAuth/JWT authorization, ownership-aware access controls, certified-table ranking, and auditable caller-role logging; scaled adoption to **120+ Data Platform engineers** and approximately **5,000 daily calls**.
  - **Frost**: Built an agentic Hive-to-Iceberg migration platform that discovers lineage, ownership, write patterns, and PII/SOX classifications; generates size-aware migration configurations; and validates changes in isolated development environments. Embedded mandatory human approval gates, CI/CD promotion controls, durable workflow state, retries, and audit records across GitLab, Jira, Slack, and Confluence.
    - **Governed Migration at Scale**: Migrated **450 tables across 40 workflows** with no production incidents, while **75% of non-critical tables required no engineering tuning** after mandatory YAML review; delivered by a **3-engineer team**, versus approximately **10 engineers** for the otherwise manual programme.
  - **Minion AI Incident Response**: Productionised Slack incident-history and Confluence runbook-retrieval MCP servers for an AI incident assistant, enforcing read allowlists, least-privilege scopes, and human-gated writes; deployed with Bazel, Terraform, and Kubernetes and adopted by **20+ teams and 150+ engineers**.
  - **GDPR Data Deletion**: Led the redesign of GDPR deletion across a **100 PB data lake**, introducing configurable compliance reporting and virtual-identifier rotation to immediately disassociate customer records; reduced end-to-end fulfilment from **90+ days to 5 days (94%)**.
  - **GDPR Data Download**: Re-architected GDPR data-download processing from MapReduce to Spark, scaling intake from **20 to 1,000+ requests/day** while limiting daily physical reads to approximately **200 TB** across a **20 PB source footprint**; partnered with legal analysts to make new compliance datasets configurable within **1–2 days**.
- **Groupon** — Bangalore, IN
  Software Development Engineer - III — *Oct 2020 – August 2024*
  - **Personalization Service**: Designed and built a Java/Dropwizard batch and real-time personalization service supporting 40M emails and 60M push notifications daily, with 5 ms p95 end-to-end latency.
  - **Consumer 360 Data Store**: Built a 200 TB+ Consumer 360 data store with 400+ behavioral, demographic, and predictive features used by Marketing Engineering.

- **Dataproc Migration**: Led the architecture and phased cutover of 2,500 Spark jobs from on-premises infrastructure to Google Cloud Dataproc; designed bidirectional table synchronization to preserve end-to-end marketing flows during per-job cutovers, then retired the synchronization layer after each channel migration without interrupting revenue paths.
- **Reusable Datastore Libraries**: Architected bulk-CRUD libraries for Amazon Keyspaces and Cloud Bigtable using batching and parallel execution to improve upsert and delete throughput; adopted as the standard access path by 10+ teams.
- **Keyspaces-to-Bigtable Migration**: Led the migration of a high-throughput service datastore from Amazon Keyspaces (Apache Cassandra) to Cloud Bigtable with no customer-visible or revenue-impacting downtime, cutting its annual run cost 90%, from approximately $2M to $200K, by replacing per-operation serverless pricing with a provisioned cluster.
- **Data Lifecycle Management**: Implemented retention and lifecycle policies in the data warehouse, reducing retained data volume by 80% and annual storage cost by approximately $500K.
- **Data Quality Framework**: Designed a framework with record-, column-, and metric-level checks and alerting across the data warehouse, reducing data-corruption incidents tracked in Jira by 75% over two quarters.

- **AtkinsRéalis** Bangalore, IN
  Software Development Engineer - II *Feb 2019 – Sep 2020*
  - Built image and GIS data pipelines on Theia, a decoupled big-data and machine-learning platform on Databricks, and optimized PySpark workloads using Spark SQL and UDFs.
  - Built serverless ETL applications that moved MongoDB data into Azure SQL Server for the Assist 2.0 workflow product.
- **Eatclub** Bangalore, IN
  Software Development Engineer - I *Jun 2018 – Jan 2019*
  - Built backend APIs for vendor onboarding and inventory tracking, including QR-code integrations with Box8's e-commerce platform.

Skills

- **Languages**: Python, Scala, Java, SQL
- **Data & Distributed Systems**: Apache Spark, Trino, Kafka, Iceberg, Hive, Hadoop, Databricks, Airflow
- **AI Data & Governance**: LLM data pipelines, synthetic data, retrieval-augmented generation, MCP, DataHub, PII classification, provenance, lineage, policy-enforced access
- **Infrastructure & Security**: Terraform, Kubernetes, Bazel, OAuth, JWT
- **Cloud & Datastores**: GCP (Dataproc, Bigtable), AWS (Keyspaces), Cassandra, PostgreSQL, DynamoDB
- **Backend**: Dropwizard (Java), FastMCP / MCP (Python)

Open Source Contributions

- **Hardwood**: Contributed to Hardwood, a minimal-dependency Java implementation of Apache Parquet: authored dictionary-space predicate evaluation (accepted, landing next release), evaluating a predicate once per column-chunk dictionary and answering rows by index comparison instead of decoding values; also restored Bloom filter pruning on `FLOAT`/`DOUBLE` columns and packed the `dive` TUI preview by content width.
- **Koalas (Databricks)**: Contributed merged DataFrame API implementations (replace, cache, to_records, to_excel) to Koalas, later merged into **Apache Spark** as the pandas API on Spark.

Education

- **Indian Institute of Technology (Indian School of Mines), Dhanbad** Dhanbad, India
  Bachelor of Technology; Minor Degree in Computer Science and Engineering; CGPA: 8.1/10 *Jul 2014 – Mar 2018*
- **The Chintels School** Kanpur, India
  XII - ICSE: 94% (Mar 2013) • X - ICSE: 93.8% (Mar 2011)

Internships

- SNC-Lavalin Nuclear Solutions (2017) • AtkinsGlobal Digital Incubator (2016) • Saint Petersburg State University, Global Citizen Fellow (2015)