

Praveen T *Data Engineer*

✉ praveenbenakannanavar@outlook.com ☎ 9535055419 📍 Bangalore

in linkedin.com/in/praveen-benakannanavar 🌐 github.com/PraveenBen

Profile

Data Engineer with 3+ years building production data platforms across streaming, batch, and BI. Hands-on with Apache Spark (Dataproc/Databricks), Apache Beam, Kafka, Pub/Sub, ClickHouse, Druid, and SQL Server. Delivered real-time stateful pipelines at Flipkart scale, near real-time CDC for renewable energy asset data, Spark cost/performance optimization, and end-to-end BI infrastructure on production data. Experienced with Azure data services, Python, and LLM-assisted development, including RLHF contributions to Google's Gemini across text, image, and audio.

Skills

Programming Languages: SQL, Shell Scripting, Python, Java

Data Platforms: Databricks, Apache Spark (Dataproc), Apache Beam, Kafka, Pub/Sub, Delta Lake, ETL/ELT, Data Warehousing, Dimensional Modeling

Databases: MSSQL, MySQL, PostgreSQL, ClickHouse, Druid

BI & Reporting: Power BI, Apache Superset, semantic/data modeling, self-service reporting, Advanced Excel (Power Query, pivot tables, lookups)

Cloud: Azure (Data Factory, Synapse, SQL), GCP (certified), AWS

AI/ML: LLM evaluation, RLHF, AI-assisted development, Vertex AI, AWS Bedrock, scikit-learn

Tools: GitHub, Airflow, Azkaban, Kubernetes, Grafana, REST APIs

Professional Experience

DataSturdy Consulting private limited

01/2025 – 09/2026 | Bengaluru, India

Associate Data Engineer

Near Real-Time CDC Platform — Solar Manufacturing & Testing Data

At DataSturdy (Tata Power Mumbai)

- Architected and delivered an end-to-end, cross-cloud near real-time CDC platform streaming **24 MSSQL tables** covering every phase of **solar panel manufacturing and testing**, from source capture through to analytics-ready storage consumed by downstream **Databricks workloads**.
- Provisioned and configured the full high-availability stack across 19 VMs — 2-node Debezium, 3-node Kafka, 5-node **Flink**, and 9-node **ClickHouse** — standing up the complete topology and end-to-end source-to-sink connectivity in **3 days**; formally **recognized by project leadership** for the turnaround.
- Designed the pipeline for exactly-once semantics, fault tolerance, and replay, ensuring no loss or duplication of stage-level test records across the manufacturing line.
- Implemented security controls across the full path — encrypted cross-cloud transport over **VPN**, restricted network routes, and least-privilege access to source and sink systems — meeting client **data governance** requirements for production manufacturing data.
- Built **Grafana** dashboards for **real-time pipeline health** and stage-level throughput monitoring, giving operations visibility into data freshness, lag, and failures.
- Provided technical guidance to a junior engineer on a companion **IoT subsystem** for **remote device power control with battery-state monitoring**.

Voice to Trade , AI-driven Bajaj Broking Project

At DataSturdy (Bajaj Broking)

- Architected an AI-driven pipeline (Voice2Trade) for Bajaj Broking to extract structured trade data from customer call recordings using **speech-to-text** and **LLM-based extraction**.
- Designed a daily automated verification workflow to reconcile extracted trade data against executed trades, ensuring customer consent compliance.
- Built the solution in adherence to SEBI regulatory requirements for trade execution and consent verification.
- Leveraged Gemini 3.5 Flash and Mistral 658B models for audio transcription and structured data extraction at scale.

Near Real-time ETL pipeline

At DataSturdy(Flipkart Ads Project)

- Built a low-latency streaming pipeline using **Apache Beam (Java SDK)** to process millions of clicks and view events from **Google Pub/Sub**. Implemented **stateful deduplication** with **RocksDB** to ensure data idempotency across Flipkart's three marketplaces (Flipkart, Hyperlocal, Grocery).
- Performed sub-millisecond lookups against **Aerospike** cache for **real-time data enrichment**. Augmented raw event streams with critical metadata before sinking to **Kafka** for downstream app.
- Deployed pipeline infrastructure on **Google Kubernetes Engine (GKE)** with **CI/CD** automation. Achieved 99.9% uptime and dynamic **auto-scaling** to handle 21x traffic spikes during peak sales events.

BI Infrastructure

At DataSturdy(Flipkart Ads Project)

- Designed and deployed a complete BI stack on Apache Superset over production data, including semantic layer definitions, parameterized dashboards, and self-service datasets for non-technical business users. Mentored junior engineers on dashboard and data model design. Program concluded following a management decision to deprioritize the initiative.

Data Sanity

At DataSturdy (Flipkart Ads Project)

- Designed and deployed a cross-platform **Data Sanity Framework** to validate data consistency across **MySQL**, **ClickHouse**, **Apache Druid**, and **GCS**. Built automated checks to ensure data integrity across multiple storage systems.
- Daily reconciliations with **Spark** ensured **financial** data quality, reducing anomalies before month-end close.
- Automated daily reporting using **Python** and **SMTP**. Generated and emailed HTML-formatted validity reports to downstream stakeholders every morning, eliminating manual validation entirely.

Pre BBD NFR Scale Testing

At DataSturdy(Flipkart Ads Project)

- Benchmarked pipeline performance against historical peak-load baselines, tuning executor memory, shuffle partitions, and mapper/reducer counts to cut compute cost while sustaining 21x traffic. Produced capacity and cost projections used for pre-event planning, with workflows orchestrated via Azkaban.
- Built and operated production Spark workloads on Google Dataproc (cluster sizing, job tuning, scheduled orchestration); same execution model as Databricks clusters.

Infosys 8 Apps optimisation

At DataSturdy (Infosys Project)

- Worked on **database performance optimization** for critical enterprise applications (**ALCON** and **OMS**). Targeted high-latency legacy **Stored Procedures** that were causing bottlenecks in production systems.
- Implemented strategic **indexing** and **query restructuring** to improve performance. Achieved a **25% average reduction** in query execution time, with some queries improving dramatically (e.g., 15 minutes down to 3 minutes).

- Successfully deployed **6 optimized Stored Procedures** to production. Modernized legacy **MSSQL** database code while maintaining data integrity and business logic.

GenAI RLHF (Freelancing)

06/2023 – Present

At Outlie.AI

- Designed **adversarial evaluation scenarios to stress-test frontier LLM agents** operating across simulated enterprise tool environments (email, chat, version control, and project tracking systems), constructing multi-hop tasks that required the model to locate, cross-reference, and verify information distributed across several systems rather than retrieve it directly.
- **Authored and ingested synthetic datasets into evaluation environments using SQL**, seeding realistic distractors and conflicting records to probe model reliability under ambiguity and surface failure modes in multi-step agentic reasoning.
- Applied the same rigor to production work, using **LLM-assisted development** for pipeline scaffolding, SQL generation, and test authoring.
- Executed **Reinforcement Learning from Human Feedback (RLHF)** workflows to fine-tune foundational LLMs including Google's Gemini, improving instruction following and constraint satisfaction across text, image, and audio modalities.
- Performed factual verification and task-specific evaluations to assess model reasoning, correcting logic errors and authoring high-quality ground-truth training data to improve model accuracy and reliability.

AffordPlan

05/2022 – 12/2022 | Bengaluru, India

Financial Counselor

- Maintained and enhanced **Apache Superset dashboards for 11+ hospital** partners under the 'Affordplan Swasth' initiative, delivering regular reports to the sales team to support partner.
- Wrote custom **SQL queries** to build **hospital-specific charts** and proactively resolved broken queries to ensure uninterrupted reporting, reducing dependency on escalations.

Education

Master of Computer Application

02/2023 – 12/2024 | Bengaluru, India

Christ Academy Institute for Advanced Studies

Bachelor of Computer Application

2018 – 2021 | Bengaluru, India

VEIT Degree College

Certificates

Google Associate Cloud Engineer 

GitHub Foundational 

Machine Learning — GIT, Hubli

TCS National Qualifier Exam — Scored 72%