

Koustubh Singh

+91 9309384317 | eharshit.dev@gmail.com | [in](#) eharshit | [G](#) eharshit
Pune - 411035, India

EDUCATION

• D Y PATIL INTERNATIONAL UNIVERSITY

B.Tech Computer Science Engineering (Data Science)

Sep 2022 – May 2026

Pune, India

SKILLS

- **Programming Languages:** Python, SQL, Golang
- **Data Engineering:** PySpark, Kafka, Hadoop, Hive, Delta Lake, Databricks
- **Databases:** MySQL, PostgreSQL, MongoDB
- **Cloud & Analytics:** AWS (S3, Glue, Athena, IAM), Power BI, Excel
- **Tools & Libraries:** Docker, Git, Airbyte, Pandas, NumPy

WORK EXPERIENCE

• LawLevelUp

Data Analyst Intern

May 2025 – Aug 2025

Ghaziabad, Uttar Pradesh

- Developed a KPI dashboard in **Power BI**, integrating real-time and predictive analytics to forecast trends, **saving more than 5 hours per week** in reporting.
- Migrated Excel-based reports to **Power BI**, enhancing accessibility and enabling interactive data exploration.
- Analyzed learning data from **2K+ enrollments** to identify the most engaging courses and recommend improvements that boosted learner participation.

PROJECTS

• PuneFlow — City Traffic Streaming Pipeline

[GitHub](#)

- Built a real-time streaming pipeline ingesting vehicle telemetry via **Apache Kafka** through Medallion Architecture (Bronze→Silver→Gold) using **Spark Structured Streaming** and **Delta Lake**.
- Designed a data quality layer handling **9 failure modes** including schema drift, duplicates and late events using watermarking and validation flags.
- Implemented Star Schema exposed via Hive Metastore to **PowerBI** dashboard tracking live congestion, speed bands and peak hour traffic across Pune city zones.

• VendorIQ — ETL & Vendor Performance Dashboard

[GitHub](#)

- Engineered ETL pipeline using Python, Pandas, SQLite processing **500K+ transactions** from 5 vendor data sources
- Built Power BI dashboard with **15+ KPIs** tracking vendor profitability, margins, and inventory health
- Identified **15% margin improvement** opportunities through vendor performance and pricing optimization analysis

• Real-Time Insurance Claims Pipeline

[GitHub](#)

- Built end-to-end pipeline processing **10K+ insurance claims** from MongoDB to **Snowflake** via Airbyte, automating daily data ingestion and loading.
- Transformed raw data using **dbt** into clean, aggregated, OLAP-ready datasets with data quality tests and incremental models for optimized processing.
- Identified **5.2% fraud rate** via Power BI, revealing Auto claims had **3x higher risk** above \$10K.

ACHIEVEMENTS & CERTIFICATIONS

• Best Innovation Award

Smart India Hackathon, Sep 2024

• Data Engineering Professional Certificate

DeepLearning.AI & AWS