DHARMESH KASHYAP

+91-89282-10971 • dharmeshkashyap46@gmail.com • linkedin.com/in/dharmesh-kashyap/ • github.com/dharmesh-kashyap • dharmeshkashyap.work

SUMMARY

Data Engineer with 2+ years of production experience delivering pipelines for NABARD and ICICI Prudential. Stack: Python, PySpark, Hadoop, Polars, DuckDB, SQL, PostgreSQL, AWS. Built PySpark/Hadoop ETL standardizing 5 NSO/MoSPI surveys across 3 rounds into a single long-format schema, a staging-to-dimension/fact PostgreSQL warehouse over 5.2M+ records, and extraction pipelines for 670K+ GeoJSON village boundaries and 2,500+ government scheme PDFs.

TECHNICAL SKILLS

Languages: Python, SQL

Big Data & Processing: PySpark, Hadoop, Polars, DuckDB, Parquet, ETL Pipelines, Data Validation

Databases & Modeling: PostgreSQL, MySQL, Dimensional Modeling (staging, dimension/fact), SQL Query Optimization

Backend & Cloud: Flask, AWS

Analytics & BI: Tableau, AIV Hub, KPI Reporting, Dashboard Development

PROFESSIONAL EXPERIENCE

Software Developer (Data Engineer) **Feb 2025 – Present**

Seamless Digital Automations, Mumbai, India

- **Survey ETL (NABARD):** Built distributed pipelines in **PySpark/Hadoop** to process raw NSO/MoSPI microdata for 5 All-India surveys (HCES, ASUSE, AIDIS, SAS, Agriculture Census) across 3 rounds, with **Polars** and **DuckDB** handling downstream transformation and validation.
- Output lands in one standardized long-format schema that BI tools consume directly, with **Parquet** recommended over CSV for delivery.
- Packaged the HCES pipeline as a modular Python ETL package with a full test suite, so the transformation logic is repeatable and verifiable.
- Built a DDI enrichment dictionary of **298 harmonized variables** across 3 HCES rounds, keeping variable definitions consistent round to round.
- **Finance warehouse (ICICI Prudential):** Designed the PostgreSQL schema for **5.2M+ records**: staging tables promoted into dimension/fact tables through a multi-step pipeline, with query optimization for analytics.
- **Extraction pipelines:** Delivered ETL for **670K+ Pan-India GeoJSON village boundaries** for NABARD geospatial reporting; built extraction with validation checks for **2,500+ government scheme PDFs**; automated a recurring pipeline consolidating **60+ mutual fund instruments** into one schema.
- **BI & infrastructure:** Built KPI dashboards with drill-downs, deployed AIV Hub on AWS-hosted servers (setup, configuration, monitoring), and documented network/firewall requirements after diagnosing a cross-subnet routing failure.

PROJECTS

Data Processor: Automated File-to-Database Platform 2025 – 2026

Flask + PostgreSQL app for finance data upload automation for ICICI Prudential

- Ingests Admin Dump and Location Vacation Excel files into staging tables, validates them, and promotes them into a dimension/fact warehouse through a multi-step pipeline.
- Added scheduler visibility, refresh progress, conflict reporting, and a notification tray; audited and fixed duplicate-promotion and staging-filter bugs.

AI-Powered News Intelligence Platform 2026

News scraping, AI enrichment, and chat platform built for NABARD

- Built a multi-source scraper that extracts, cleans, and structures articles into a database feeding a live dashboard.
- Added AI enrichment for classification, summarization, and search, plus an interactive commodities ticker.

EDUCATION

Master of Computer Applications (MCA) 2023 – 2025

Manipal Institute of Technology, Manipal Academy of Higher Education

Coursework: Data Analytics, Database Management Systems, Cloud Computing

Bachelor of Science in Information Technology (B.Sc. IT) 2019 – 2022

Ramniranjan Jhunjhunwala College, University of Mumbai

CERTIFICATIONS

Modern Application Development with Python on AWS • Applied Data Science with Python