

# SONAM KUMARI

Jamshedpur, Jharkhand — +91 8434047566 — er.sonam2712kumari@gmail.com  
linkedin.com/in/sonamkumari27 — github.com/SonamKumari1227 — Personal Website

## Summary

---

Data Engineer / Python Developer with 2+ years of experience building scalable ETL pipelines, web ingestion systems, and automation workflows. Designed and maintained 2000+ production-grade web scrapers supporting full and incremental updates. Strong in Python, SQL (PostgreSQL), Docker, and Azure Data Factory, with hands-on experience in PySpark-based processing and data warehousing.

## Education

---

### Ajay Kumar Garg Engineering College, Ghaziabad, India

Bachelor of Technology in Computer Science and Engineering (Data Science)  
CGPA: 7.92/10

2021 – 2025

## Experience

---

### Nimbus Maps

Junior Data Engineer — Remote

Feb 2024 – Present

Jun 2025 – Present

- Designed and maintained end-to-end Azure Data Factory pipelines supporting full and incremental ingestion for large-scale web data workflows.
- Built and optimized high-volume web scraping systems handling millions of records across multiple sources.
- Containerized ingestion pipelines using Docker, enabling consistent deployments across environments.
- Collaborated using GitHub-based CI/CD workflows to manage versioned data pipelines.

Data Engineer Intern — Remote

Feb 2024 – May 2025

- Extracted real estate data from 2000+ commercial property websites using production-grade web scrapers.
- Improved scraper efficiency by 25% through debugging, optimization, and performance tuning.
- Built reusable scraping libraries, reducing new scraper development time by 40%.

### Cloud Computing Cell, AKGEC

Full Stack Web Developer — AWS Cloud

Oct 2022 – March 2025

- Worked on backend-focused web applications with cloud deployment using AWS services.
- Implemented hosting, deployment, and server-side integration for scalable applications.

## Projects

---

### Real Estate Analytics Pipeline (PySpark, Bronze/Silver/Gold)

GitHub

- Built an end-to-end real estate analytics pipeline in PySpark using Bronze → Silver → Gold architecture.
- Converted raw CSV data into Parquet, performed cleaning, deduplication, and datatype standardization.
- Generated KPI tables for city-level insights, portfolio analysis, property distribution, and rental yield metrics.

### Retail Data Pipeline (Python, Pandas, SQL (PostgreSQL))

GitHub

- Built an end-to-end ETL pipeline to generate, clean, transform, and load retail sales data into PostgreSQL.
- Implemented data cleaning and validation using Pandas (handling nulls, duplicates, schema consistency, and type corrections).
- Designed a star schema model including fact\_sales and dimension tables for customers and products.
- Developed SQL analytics queries and views for revenue trends, top customers, category-level revenue, and monthly growth insights.

## Certifications

---

- HackerRank Certified – SQL (Advanced) (2026)
- DP-203: Data Engineering on Microsoft Azure – Udemy

## Technical Skills

---

**Languages:** Python, SQL, JavaScript

**Backend & APIs:** FastAPI, REST APIs, Microservices Architecture

**Data Engineering:** PySpark, Apache Spark, Azure Data Factory (ADF), Azure Databricks, Data Warehousing

**Web Scraping:** Scrapy, Selenium, BeautifulSoup, lxml, Proxies, CAPTCHA Handling, Anti-Bot Handling

**Databases:** PostgreSQL, MongoDB

**Tools:** Docker, GitHub, Azure DevOps, Pandas, Power BI