Yuvraj Singh

yuvraj.singh212@gmail.com 7354460006 LinkedIn GitHub

Education

- **B.tech in Computer Science** *(e-commerce technology)*, VIT, Bhopal (May, 2026) **8.09 CGPA**

Technical Skills

- **Languages & Databases:** Python, PostgreSQL, Bash
- **Data Libraries & Tools:** Pandas, PySpark, Streamlit, RegEx, Airflow, Azure VM, Azure Redis, Azure Data Lake, Docker
- **Technical Proficiencies:** Web Crawling, Web Scraping, ETL Pipelines, API Integration, Automation Scripts

Projects

B2B Cloud Data Pipeline and EDA

[ETL, EDA, Azure VM, Apache Airflow, Azure Redis, ADLS, PostgreSQL, Web-hook]

- **Situation:** Needed to acquire a 350k+ record B2B dataset for supply chain analysis, but manual collection was unfeasible due to the source's scale and anti-bot measures.
- **Task:** To design, build, and deploy a scalable, fully automated, cloud-native ETL pipeline with a user-friendly trigger mechanism.
- **Action:** Architected a complete data platform on **Microsoft Azure**, orchestrating the ETL workflow with **Apache Airflow**, containerized with **Docker**. Engineered a resilient API scraper and feature extraction pipeline that persists data through a medallion architecture on **Azure Data Lake Storage (ADLS)**, using **Azure PostgreSQL** and **Redis** as robust, managed backends.
- **Result:** Successfully deployed an end-to-end data application, complete with a **Streamlit** UI for triggering new pipeline runs via the Airflow REST API. The automated system produced a clean, analysis-ready dataset that enabled key EDA insights into regional market concentration and supplier trends.

Shopinion

[Python, Apache Airflow, Docker, Selenium-Stealth, Pandas, Azure PostgreSQL, ETL]

- **Situation:** A project to train a context-aware BERT sentiment model required a **large-scale dataset** of over 100,000+ product reviews, but manual collection was unfeasible due to Flipkart's robust **anti-scraping measures**.
- **Task:** To engineer a **fully automated, end-to-end ETL pipeline** that autonomously handled the entire data lifecycle—from stealthy web scraping to structured storage and provided a simple interface for others to use.
- **Action:** Orchestrated the entire workflow using **Apache Airflow** and **Docker**, designing DAGs for parallel execution. Engineered a resilient scraper with **Selenium-Stealth** to bypass anti-bot measures, **automatically extracting over 100,000+ clean, unique reviews**. Implemented an intelligent loading module that performed a pre-emptive check to ensure **idempotent writes** and prevent data duplication.
- **Result:** Deployed a robust, **self-service data collection platform**. The parallel processing mode demonstrated a **20-22% improvement in execution time** over sequential scraping and enabled others to self-serve the creation of large-scale datasets.

Outbreak Tracker

[Python, Pandas, PyFaker, Streamlit, Scikit-learn]

- **Situation:** Inspired by a local jaundice outbreak, recognized that standard **diagnostics failed** without awareness of real-time, **regional case counts**.
- **Task:** To overcome the complete **unavailability of a suitable public dataset** by **engineering a custom dataset** from scratch for a proof-of-concept.
- **Action:** Built the **100k+ record dataset** by **merging real state-wise statistics** with **synthetically generated** patient profiles from PyFaker, then developed a **Streamlit app** that applied **location-based rules** to predictions.
- **Result:** The deployed app **cut critical misclassifications by 22%** and **boosted overall diagnostic accuracy by 5-10%**.

Advanced Data Cleaning and Feature Engineering

[Python, Pandas, RegEx, Data Cleaning, ETL, Feature Engineering]

- **Scripted a data validation pipeline** using Pandas and RegEx, **removing 32% duplicate records** and **standardizing inconsistent text** into structured columns to create a validated dataset.
- **Eliminated over 20% of critical data gaps** by **engineering a new, high-fidelity 'Type' column** through a logical pipeline that classified content based on director tags, runtime data, and genre-based **keyword matching**.

Extracurricular Activities

- **SGFI National-level Basketball Player**; represented at multiple regional and inter-school tournaments
- Member of **university sports council**, contributing to **planning and execution** of intra-college leagues

Certifications

- **Data Engineer Associate** (Data Camp, Feb 2025)
  Learned ETL workflows, SQL for data modeling, and database design using PostgreSQL and Snowflake.