Harsh Indoria

+91 8824356811 | harsh.ind.coder@gmail.com | linkedin.com/in/harsh-indoria | github.com/harshindcoder

Education

Integrated MSc in Applied Mathematics, IIT Roorkee | CGPA: 9.09/10 Aug 2019 – June 2024

- Relevant Coursework: DSA, DBMS, Data Mining for Business Intelligence, Probability & Statistics

Technologies

Programming & Scripting: Python, SQL, Bash
Big Data & Workflow Orchestration: Apache Spark, Airflow
Databases & Storage: MySQL, MongoDB
Cloud Platforms: Google Cloud Platform
Python Libraries: NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Regex, PySpark
Version Control & DevOps: dbt, Git, GitHub, Docker

Experience

Data Tech, Kotak Mahindra Bank Ltd – Mumbai, IND May 2023 – July 2023

Technologies: ElementTree XML API, Python, PySpark, Regex, Bash, AWS.

- Worked with and tested multiple document parsing libraries in Python, including ElementTree, BeautifulSoup, and lxml for data transformation improvement.
- Collaborated with system administrators to analyze Salesforce CRM data and with the Data Warehouse team on data modeling, improving accessibility and usability of unstructured data.
- Improved document parsing performance by 5–10x and enhanced data visibility by transforming unstructured data into structured formats, enabling ingestion from AWS S3 into AWS Redshift using PySpark.

Research Assistant, IIT Roorkee – Roorkee, IND June 2022 – Jan 2023

- Worked on developing and improving mathematical models used in industries under the guidance of Dr. Madhu Jain, resulting in a research paper published in RAIRO-Operations Research, January 2025.
- Presented the research findings at the USTM-AIMT Summer International Conference 2022 in Meghalaya, India.

Projects

GCP Data Engineering Pipelines [Link] Aug 2025 – Present

Technologies: GCP (BigQuery, CloudSQL, GCS), SQL, Python, Bash

- Developed an end-to-end ELT pipeline on GCP (CloudSQL $\rightarrow$ GCS $\rightarrow$ BigQuery) to enable scalable analytics.
- Documented and version-controlled the entire process, and continue to expand it with additional tools.
- Automated data ingestion, transformation, and schema design with SQL and Python for analytics readiness.

Customer Segmentation Using K-means Clustering [Link] Dec 2024

- This marketing analytics project uses RFM (Recency, Frequency, Monetary) features for customer classification, inspired by the online retail mining paper.
- The RFM model helps segment customers, identify high-value ones, and optimize marketing strategies.

Certifications

Google Advance Data Analytics, Google Data Analytics, Python for Data Science and Machine Learning Bootcamp, Probability and Statistics for Business and Data Science

Positions of Responsibility and Extracurricular Activities

NGO Relations and Sponsorship Team Member, NSS IIT Roorkee – Roorkee, IND July 2019 – June 2022