SHIVRAJ ROMAN

Data Engineer

+91 8652604789 | shivrajroman@gmail.com | Mumbai, IN | linkedin.com/in/shivrajroman

CAREER SUMMARY

Post graduate in Big Data Analytics from CDAC Mumbai with a strong foundation in Python, SQL, Cloud Computing, PySpark Machine Learning, and Data Visualization. Hands on experience in real time projects using modern data tools such as Python, PySpark, SQL and Azure Databricks to develop scalable big data solutions.

Skilled in planning and contributing to backlog refinement, gap findings, process optimization and cross-team co-ordination to deliver high-quality solutions on schedule.

SKILLS

Languages: SQL, Python, R

Big Data Frameworks: Spark, PySpark, SparkSQL, Hadoop

Cloud Technologies: Azure Data Factory, Azure Data Lake, Databricks, AWS

Data Visualization: Power BI and Tableau

Other Tools: GitHub, Microsoft Excel and Power Point

PROJECTS

FRAUD DETECTION PIPELINE | Azure Databricks, Python, ML, Power BI, Streamlit | CDAC

Built a scalable fraud detection pipeline with a complete ETL process using Azure Databricks and Delta Lake. Handled data ingestion, transformation, and ML-based classification of banking transactions. Delivered real-time fraud insights via Power BI dashboards and a Streamlit UI.

- Executed ETL on 1M+ records using Spark SQL; trained ML models (Logistic Regression, KNN) with strong accuracy.
- Automated pipeline with Python; applied rule-based logic via Delta metadata for dynamic processing.

CIPLA STOCK ANALYSIS | PySpark, MatplotLib & Power BI | CDAC

Performed stock data analysis for Cipla using a PySpark-based ETL pipeline to process and analyze historical trends. Applied transformations and time-series logic to extract insights on stock performance. Visualized patterns using Matplotlib and summarized findings in Power BI dashboards.

- Processed large CSV datasets with PySpark; calculated moving averages, price trends, and volume analytics.
- Built interactive dashboards in Power BI for executive-level visualization and trend interpretation.

IMDB CASE STUDY | Python, Pandas & Power BI | CDAC

Scraped and analyzed IMDb data (2000–2019) to identify trends across genres, budgets, and profitability. Performed ETL and statistical analysis to uncover patterns in production spending and returns. Used Pandas for data wrangling and visualized insights on genre-wise performance metrics.

- Extracted and cleaned raw IMDb data using Python scripts and API; filtered out incomplete entries.
- Found that genres like History, Mystery, and Horror offer high ROI, while Adventure shows inconsistent profitability.

EDUCATION

PG DIPLOMA | Big Data Analytics | CDAC-Mumbai — Aug 2024 – Feb 2025

BACHELOR'S DEGREE | Engineering | Pune University — Jun 2016 – May 2020

CERTIFICATION

Python, SQL, Java (HackerRank), Databricks Fundamentals, Microsoft Excel (SimpliLearn)

INTERESTS

Travelling | Abstract photography | Social Services | Exploring | Trading | Playing Video Games