

LAKSHAY MITTAL

W-18, Malimandir, Puldarwaja, Sheopur, MP 476337

+91-6266014897 ✉ lakshay.mi05@gmail.com in [linkedin.com/in/lakshay-mittal](https://www.linkedin.com/in/lakshay-mittal) 🐙 github.com/sicario5

Summary

Data Engineer skilled in building **ETL pipelines**, automating workflows, and designing scalable data systems. Strong expertise in **Python, SQL, AWS, PySpark, Kafka**, and cloud-based architectures for data-driven insights and cost optimization.

Education

Maulana Azad National Institute of Technology, Bhopal

B.Tech., Electrical Engineering

Jul 2018 – May 2022

GPA: 7.17

Experience

Data Engineer

AU Small Finance Bank

June 2022 - November 2023

Jaipur, Rajasthan

- Built a robust **ETL framework** with **AWS EMR, S3, Redshift, Python, and SQL** for efficient data ingestion and transformation.
- Automated workflows with optimized **Apache Airflow** DAGs, improving efficiency and reliability.
- Designed and optimized **data models** (de-normalized / aggregated tables) for dashboards and **KPIs**.
- Processed **3M+ daily transactions** (5-year history) with **PySpark on AWS EMR**, implementing incremental processing to achieve a **60% cost reduction**.
- Deployed **PySpark** solutions on AWS EMR to identify potential NPA customers and automated their daily email alerts for using **AWS Lambda**.
- Developed and optimized SQL and PySpark queries for efficient data retrieval and analysis, achieving a **20% reduction in EMR cluster costs** through effective resource management.

Projects

Real-Time Flight Data Pipeline [Github](#)

April 2025

- Developed a real-time flight data pipeline using Selenium, Pandas, and Kafka, automating Kayak flight scraping with 15s throttling and streaming data to S3 with efficient date-based partitioning.
- Reduced AWS storage costs by **40% through JSON data compression** and S3 lifecycle management policies
- **Technology:** Jupyter Notebook, Pandas, Selenium, Kafka, AWS, Git-Bash

IPL Data Pipeline [Github](#)

August 2025

- Designed and implemented a Kafka-based real-time ingestion system to read multiple IPL CSV datasets, batch process rows in the consumer for optimized S3 writes, and store them in Amazon S3.
- Performed transformations in Databricks using PySpark and SparkSQL for cleaning, joining, and enrichment, followed by data visualization for match insights.
- **Technology:** AWS, Kafka, DataBricks, PySpark, SparkSQL, matplotlib, seaborn, Git Bash

Skills

Languages:	Python, SQL, SparkSQL
Big Data:	PySpark, Spark, Kafka, Selenium, Pandas
Cloud:	AWS (Athena, S3, Glue, EMR, Lambda, Redshift), Azure, GCP
ETL Tools:	SSIS, Byte App (Accenture), ETL Pipelines
Database:	MySQL, MongoDB, Data Warehouses, Data Modeling
Tools:	Databricks, Tableau, Power BI, Git, MS Office
Core CS:	Data Structures, Algorithms, DBMS, OS, OOPs

Certifications

- **Databricks Fundamentals Accreditation:** [Databricks](#)
- **4-Star Python Badge:** [HackerRank](#)
- **Supervised Machine Learning: Regression and Classification :** [Coursera](#)
- **ETL and Data Pipelines with Shell, Airflow and Kafka:** [Coursera](#)