Piyush Pandey

+91 9311562074 piyushpandey4263@gmail.com linkedin.com/in/piyush-pandey03

Education

VIT Bhopal University, Bhopal, India *Oct 2022 – Apr 2027*
Integrated M.Tech in Data Science (CGPA: 8.07)

Professional Experience

Data Analyst Intern *Nov 2025 – Jan 2026*
Ernst & Young (EY)

- Designed and deployed scalable Parquet-based data processing pipelines using PySpark, incorporating schema validation, error handling, and data quality monitoring to ensure reliable enterprise data ingestion.
- Optimized Delta Lake table architecture by improving partitioning strategies, tuning query execution plans, and restructuring storage layouts to enhance performance for analytical workloads.
- Automated end-to-end ETL workflows from Databricks to MySQL using Python connectors, implementing scheduling, logging, and recovery mechanisms to improve data synchronization reliability.

Projects

Enterprise Parquet-to-Analytics Pipeline *Nov 2025 – Jan 2026*
Databricks, PySpark, MySQL, Delta Lake

- Developed a fault-tolerant PySpark pipeline for large-scale Parquet data processing, integrating checkpointing, validation, and retry mechanisms to ensure consistent batch ingestion.
- Implemented automated schema drift detection and metadata tracking systems to handle evolving data structures and minimize long-term pipeline failures.
- Designed an integrated architecture using Delta Lake and MySQL to enable efficient storage, fast querying, and seamless integration with reporting tools.

Databricks Lakehouse (Medallion Architecture) *Oct 2025 – Nov 2025*
PySpark, Delta Lake, DLT, SQL

- Architected a scalable Medallion Architecture using Bronze, Silver, and Gold layers to organize raw, refined, and curated datasets for analytical consumption.
- Implemented Slowly Changing Dimensions using Delta Live Tables to maintain historical records, ensure data consistency, and support trend analysis.
- Orchestrated automated ETL workflows using Databricks Jobs by configuring scheduling, dependency management, and monitoring systems for continuous processing.

Customer Churn Prediction using Machine Learning *Aug 2025 – Sept 2025*
Python, Scikit-learn, XGBoost, Pandas

- Built an end-to-end machine learning pipeline including data preprocessing, feature engineering, model training, and evaluation for churn prediction.
- Tuned XGBoost and Random Forest models using cross-validation and hyperparameter optimization techniques to improve generalization performance.
- Conducted feature importance and SHAP-based interpretability analysis to identify key churn drivers and support business decision-making.

Technical Skills

Programming & Query Languages: Python, SQL, Java, C++
Data Engineering & Cloud: PySpark, Azure Databricks, Delta Lake, ADLS Gen2, Azure Data Factory
Analytics & Machine Learning: Pandas, NumPy, Scikit-learn, Feature Engineering, EDA
Visualization & Tools: Power BI, SQL Warehouses, MySQL, Git, Azure DevOps, Delta Live Tables (DLT)

Certifications

Applied Machine Learning — Coursera
Supervised and unsupervised learning, model evaluation, feature engineering, and practical ML workflows using Python and Scikit-learn.