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

Professional Experience

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

- Designed and deployed scalable Parquet data processing pipelines on Azure Databricks using PySpark, implementing schema validation and data quality checks that improved data accuracy by 25% and reduced downstream errors.
- Optimized Delta Lake table architecture on Databricks, achieving a 40% reduction in query execution time and enabling real-time analytics for stakeholders across 5+ departments.
- Automated end-to-end ETL workflows from Azure Databricks to MySQL using Python connectors, reducing manual data synchronization time by 80% and ensuring 99.9% data consistency.
- Collaborated with cross-functional teams including product managers and business analysts to translate complex requirements into actionable data models, supporting 10+ key business initiatives.

Projects

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

- Architected a production-grade Medallion Architecture (Bronze, Silver, Gold) using Databricks Auto Loader, processing 2M+ records daily with 99.5% uptime.
- Implemented SCD Type 1 and Type 2 using Delta Live Tables (DLT), preserving historical data lineage and enforcing quality rules that flagged 15% anomalous records.
- Orchestrated automated ETL workflows via Databricks Jobs and delivered BI-ready datasets through SQL Warehouses, reducing report generation time by 60%.

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

- Built a fault-tolerant PySpark pipeline to process 500GB+ of Parquet data daily with automated schema drift detection.
- Designed a hybrid architecture integrating Delta Lake with MySQL, enabling analytical and operational reporting and improving ad-hoc query performance by 35%.

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

- Developed an end-to-end ML pipeline achieving 87% churn prediction accuracy using EDA, feature engineering, and data preprocessing on 50,000+ records.
- Tuned XGBoost and Random Forest models with 5-fold cross-validation, improving performance by 12% over baseline logistic regression.
- Used SHAP and feature importance analysis to identify key churn drivers, producing insights that could reduce churn by 18%.

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, Statistical Modeling, Predictive Analytics
Visualization & Tools: Power BI, SQL Warehouses, MySQL, Git, Azure DevOps, Delta Live Tables (DLT)

Certifications

Applied Machine Learning *Coursera*
Covered supervised and unsupervised learning, model evaluation, feature engineering, and practical machine learning workflows using Python and Scikit-learn.