Piyush Pandey

piyush-pandey03 | piyushpandey4263@gmail.com | +91 9311562074

Professional Experience

Data Analyst Intern, Ernst & Young (EY) November 2025 - January 2026

- Preprocessed large-scale Parquet datasets in Azure Databricks using PySpark, ensuring schema consistency, null handling, and data quality validation.
- Designed and maintained Databricks-managed tables optimized for analytical workloads and downstream reporting.
- Implemented automated data ingestion pipelines from Databricks to MySQL using Python connectors and SQL-based transformations.
- Collaborated with cross-functional teams to translate business requirements into analytics-ready data models.

Projects

Enterprise Parquet-to-Analytics Pipeline using Databricks 2025

- Developed a scalable PySpark-based data pipeline in Azure Databricks to preprocess and validate large Parquet datasets with schema enforcement and data quality checks.
- Built Databricks-managed Delta tables and automated ingestion into MySQL using Python and SQL, enabling efficient downstream analytics.

Databricks Lakehouse with Auto Loader and Delta Live Tables 2025

- Designed and implemented a full lakehouse architecture (bronze, silver, gold layers) using Azure Databricks Auto Loader for incremental and idempotent data ingestion.
- Applied PySpark transformations in the silver layer and built gold-layer star schemas for BI consumption.
- Implemented Slowly Changing Dimensions (SCD Type 1 in PySpark and SCD Type 2 using Delta Live Tables) with data quality expectations and lineage tracking.
- Orchestrated ETL workflows using Databricks Jobs and published BI-ready datasets via SQL Warehouse.

Customer Churn Prediction using Machine Learning 2025

- Built an end-to-end machine learning pipeline using Python, Pandas, and Scikit-learn to predict customer churn from structured business data.
- Performed data cleaning, feature engineering, and exploratory data analysis to identify key churn drivers.
- Trained and evaluated multiple models including Logistic Regression, Random Forest, and XGBoost, optimizing performance using cross-validation and hyperparameter tuning.
- Achieved high predictive accuracy and interpreted results using feature importance and model explainability techniques.

Education

VIT Bhopal University Oct. 2022 - Apr. 2027

- Integrated M.Tech – Data Science (CGPA: 8.07)

Kendriya Vidyalaya No. 2 May 2021 - Apr. 2022

- Class XII (Percentage: 79%)

Skills

Languages: Python, SQL, Java, C++

Data Science: Pandas, NumPy, Scikit-learn, Feature Engineering, EDA, Machine Learning

Big Data & Cloud: PySpark, Azure Databricks, Delta Lake, Azure Data Lake Storage (ADLS Gen2), Azure Data Factory

Tools: Power BI, Git, Azure DevOps, MySQL