Python Polars Training Course

Duration: 4 days Course
Prerequisites: Knowledge of Python Programming

Day 1: Introduction & Core Foundations (6–7 hrs)

1. Course Overview & Environment Setup
    - Installing Polars, Arrow, Parquet support
    - Setting up Jupyter/VS Code
    - Sample datasets (CSV, Parquet, JSON)
2. Why Polars?
    - Pandas vs Polars vs Arrow
    - Columnar data model and Rust engine basics
3. Eager Execution API
    - Series and DataFrame basics
    - Column selection, filtering, renaming
    - Data type handling & null values
4. Expressions and Column Operations
    - pl.col, pl.lit, conditional expressions
    - String, datetime, numeric expressions
5. Labs
    - Load datasets (CSV + Parquet)
    - Perform column selections, filters, renames
    - Implement string/date transformations

Day 2: Lazy API & Transformations (6–7 hrs)

1. Lazy vs Eager Execution
    - scan_csv / scan_parquet
    - Query optimization, .collect(), .fetch(), .explain()
2. Transformations
    - select, with_columns, drop
    - Aggregations & group_by operations
    - Sorting, unique, joins

3. Window Functions & Advanced Expressions
    - Ranking, rolling windows, cumulative ops
    - Conditional aggregations
4. Performance Features
    - Predicate pushdown & projection pruning
    - Parallel execution concepts
5. Labs
    - Build a lazy query for sales analysis
    - Implement joins + window ops
    - Compare eager vs lazy performance

Day 3: Advanced Topics & Integrations (6–7 hrs)

1. Time Series & DateTime Operations
    - Resampling, shifting, time windows
    - Handling durations, offsets, calendars
2. Working with Complex Data Types
    - Structs, lists, nested data
    - Explode, flatten, and pivoting data
3. Interoperability
    - With pandas, NumPy, Arrow, PySpark
    - Writing back to CSV, Parquet, IPC
4. Streaming & Memory Management
    - Streaming CSV/JSON reading
    - Memory-efficient joins and filters
5. Labs
    - Time-series resampling of stock or taxi dataset
    - Nested JSON parsing into Polars
    - Export cleaned datasets to Parquet

Day 4: Optimization, Real-World Pipelines & Capstone (6–7 hrs)

1. Performance Tuning & Best Practices
    - Column pruning, query fusion
    - Using collect(streaming=True)
    - Benchmarking vs pandas
2. Real-World Case Studies
    - ETL pipeline with multiple joins
    - Large dataset analytics (100M+ rows demo)
3. Capstone Project (End-to-End)
    - Ingest raw sales/log data (CSV/Parquet)
    - Apply cleaning, transformations, aggregations
    - Time-series + window calculations