School Enrollment & Education Performance Analytics Platform

Project Summary

This project focuses on building a centralized analytics platform to analyze **school enrollment and education performance data** using **Python, Databricks, Airflow, and Power BI**. The solution addresses challenges such as fragmented data sources, manual reporting, limited visibility into trends, and difficulty identifying dropout risks.

The pipeline follows a **Bronze–Silver–Gold architecture**, enabling scalable ingestion, data quality validation, advanced analytics, and automated reporting.

Problem Statement

Educational data is often:

- Stored across **multiple disconnected files**
- Processed manually, making reporting **time-consuming**
- Lacking actionable insights on **enrollment trends, performance, and dropout risks**

The objective is to:

- Automate data processing
- Track enrollment and performance trends
- Enable data-driven education planning
- Provide interactive dashboards for administrators

Architecture Overview

Data Sources

- Year-wise enrollment CSV files (2020–2024)

Processing & Storage

- Databricks (PySpark & Pandas)
- Delta Lake tables (Bronze, Silver, Gold)

Orchestration

- Apache Airflow (DAG-based scheduling)

Visualization

- Power BI dashboards

Pipeline Design

1. Data Ingestion (Bronze Layer)

- Ingests enrollment data from multiple disconnected files
- Combines all years into a unified dataset
- Adds raw student performance indicators during ingestion
- Tracks source metadata for traceability

2. Data Cleaning & Validation (Silver Layer)

- Handles missing values and inconsistencies
- Standardizes categorical fields (gender, district, grades)
- Validates numeric ranges (scores, attendance, percentages)
- Ensures time-series consistency across academic years

3. Analytics & Aggregations (Gold Layer)

Creates analytics-ready Delta tables for:

- Year-wise enrollment trends
- Gender-wise and grade-level enrollment
- District-wise enrollment and performance
- Average exam scores and pass rates
- Attendance vs performance
- Learning growth and skill index trends
- Resource efficiency and remedial impact
- Learning trajectory classification of schools

Advanced Analytics & Machine Learning

To add value beyond standard reporting, the project includes:

Dropout Risk Analysis

- Engineered dropout risk score using performance and behavioral indicators
- Identifies high-risk schools and districts
- Supports early intervention planning

Learning Growth & Resource Insights

- Analyzes factors influencing learning improvement
- Evaluates remedial effectiveness
- Assesses student–teacher ratio efficiency

Visualization (Power BI)

Interactive dashboards are organized by theme:

- **Enrollment Overview**
- **Academic Performance**
- **Learning Growth & Skill Readiness**
- **Dropout Risk & Early Warning**
- **District & School Comparisons**

Dashboards enable filtering by year, district, and school to support decision-making.

Workflow Automation

- End-to-end pipeline orchestration using **Apache Airflow**
- Databricks jobs triggered for ingestion, cleaning, and analytics
- Retry logic and scheduling implemented
- Power BI refresh step represented as part of the workflow

Key Outcomes

- Centralized, automated education analytics pipeline
- Improved visibility into enrollment and performance trends
- Early identification of dropout risks
- Scalable architecture suitable for enterprise deployment

Technologies Used

- Python (Pandas, PySpark)
- Databricks Community Edition
- Delta Lake
- Apache Airflow
- Power BI

Conclusion

This project demonstrates a **real-world analytics pipeline** that combines data engineering, analytics, machine learning, and automation. It not only fulfills the stated requirements but also extends them with predictive insights and explainable analytics, making it suitable for data-driven education planning at scale.