# Student Dropout Risk & Education Retention Analytics System

An End-to-End Data Analytics Application using Python, Databricks, Apache Airflow, and Power BI


**ABSTRACT**

Educational institutions face increasing challenges in retaining students due to academic difficulties, irregular attendance, and socio-economic factors. Although large volumes of student data are collected, the lack of integrated analytics systems prevents early identification of students at risk of dropping out.

This project presents an end-to-end Student Dropout Risk & Education Retention Analytics System that integrates academic performance, attendance behavior, demographic, and socio-economic data to identify high-risk students and support early intervention strategies. The system is implemented using a scalable medallion architecture (Bronze, Silver, Gold) on Databricks, with workflow orchestration managed through Databricks Workflows and Apache Airflow.

Advanced analytics and risk indicators are visualized through Power BI dashboards, enabling educators and administrators to monitor retention trends, detect early warning signals, and plan targeted interventions. The solution demonstrates real-world data engineering, analytics, orchestration, and visualization practices in the education domain..


## 1. BUSINESS CONTEXT & PROBLEM STATEMENT

Educational departments and academic institutions collect student-related data from multiple systems such as learning management systems, attendance registers, academic records, and administrative databases. However, these datasets are often fragmented and analyzed manually, resulting in delayed identification of students who are disengaging or at risk of dropping out.

The key challenges faced by institutions include:

- Student data distributed across multiple academic and administrative systems
- Manual analysis that fails to detect early warning signs
- Lack of correlation between attendance, academic performance, and demographic factors
- Absence of actionable dashboards to support timely intervention

Due to these limitations, institutions often respond to dropout risks only after students have already disengaged. This project aims to address these challenges by building a unified analytics platform that supports proactive, data-driven retention strategies.

## 2. PROJECT OBJECTIVES

- The primary objectives of this project are:
- To design and implement an end-to-end education analytics pipeline
- To ingest, clean, and transform raw student datasets using Python and Spark
- To identify students at risk of dropping out using academic and attendance indicators
- To automate data workflows using Apache Airflow and Databricks Workflows
- To visualize retention KPIs and dropout risk insights using Power BI
- To demonstrate real-world data engineering and analytics practices

## 3. TECHNOLOGY STACK

Python, Pandas, NumPy, PySpark, Databricks, Apache Airflow, Power BI, GitHub

## 4. DATA DESCRIPTION & SCHEMA DESIGN

The system uses **five datasets** representing different aspects of student information. Approximately **5,000–6,000 records** were generated to simulate a realistic institutional dataset. Intentional inconsistencies and missing values were introduced to demonstrate data cleaning and transformation in the Silver layer.

| Dataset | Description |
|---|---|
| students_master | Student demographic and enrollment information |
| academics | Academic performance and grades |
| attendance | Attendance and participation records |
| socio_economic | Family income, internet access, employment |
| retention | Dropout and retention outcomes |

## 4.1 Dataset-wise Schema Description

### 4.1.1 Students Master Dataset

**Purpose:**
Acts as the **core dimension table**, containing stable student attributes used to join all other datasets.

**Key Columns:**

- student_id – Unique identifier for each student (Primary Key)
- gender – Student gender
- date_of_birth – Used for age-based analysis
- region – Geographic region
- school_id – School or institution identifier
- grade_level – Current academic grade
- enrollment_year – Year of enrollment
- program_type – Academic program category

**Why this dataset is important:**
It forms the **base entity** for all analytics and ensures consistent student identification across datasets.

### 4.1.2 Academic Performance Dataset

**Purpose:**
Captures student academic outcomes across semesters and academic years.

**Key Columns:**

- student_id – Foreign key referencing students
- academic_year – Year of assessment
- semester – Semester number
- avg_score – Average academic score
- failed_subjects – Number of failed subjects

**Significance:**
Academic performance is a **primary predictor of dropout risk**, especially when combined with attendance data.

### 4.1.3 Attendance Dataset

**Purpose:**
Tracks student engagement and participation.

**Key Columns:**

- student_id
- academic_year
- total_days
- days_present

**Derived Metrics (created later):**

- Attendance percentage
- Chronic absentee flag

**Significance:**
Attendance behavior often signals disengagement **before academic failure occurs**, making it a critical early warning indicator.

### 4.1.4 Socio-Economic Dataset

**Purpose:**
Represents external and demographic factors that influence retention.

**Key Columns:**

- family_income_band
- parental_education
- access_to_internet
- working_part_time
- scholarship_status

**Significance:**
Socio-economic stressors strongly influence dropout likelihood and help explain **non-academic risk factors**.

### 4.1.5 Retention Dataset

**Purpose:**
Captures the **final outcome** of student retention or dropout.

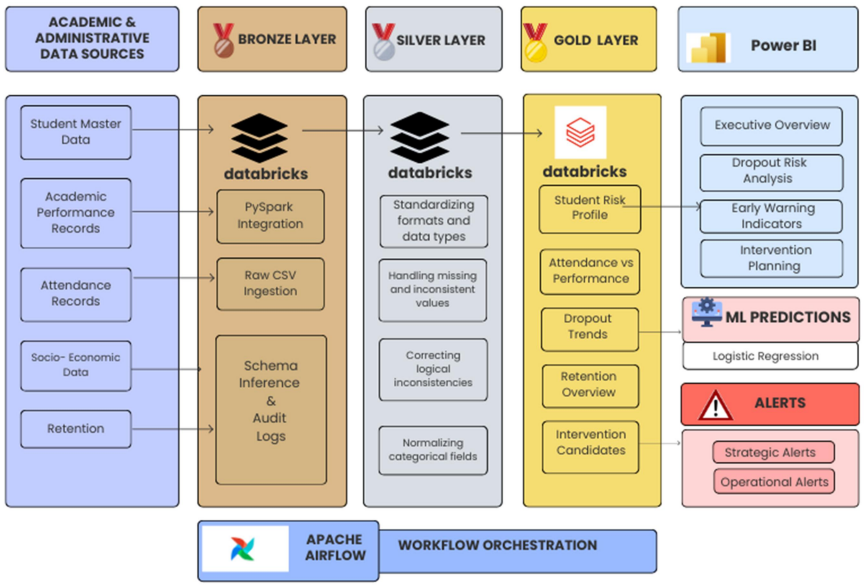
**Key Columns:**

- dropout_flag
- dropout_reason
- retention_status
- intervention_required

**Significance:**
Used for validating risk predictions and measuring the effectiveness of early warning indicators.

.

5. SYSTEM ARCHITECTURE & DATA FLOW




## 6. System Architecture & Data Flow

This section explains **exactly how data moves** through the system.

### 6.1 High-Level Architecture Overview

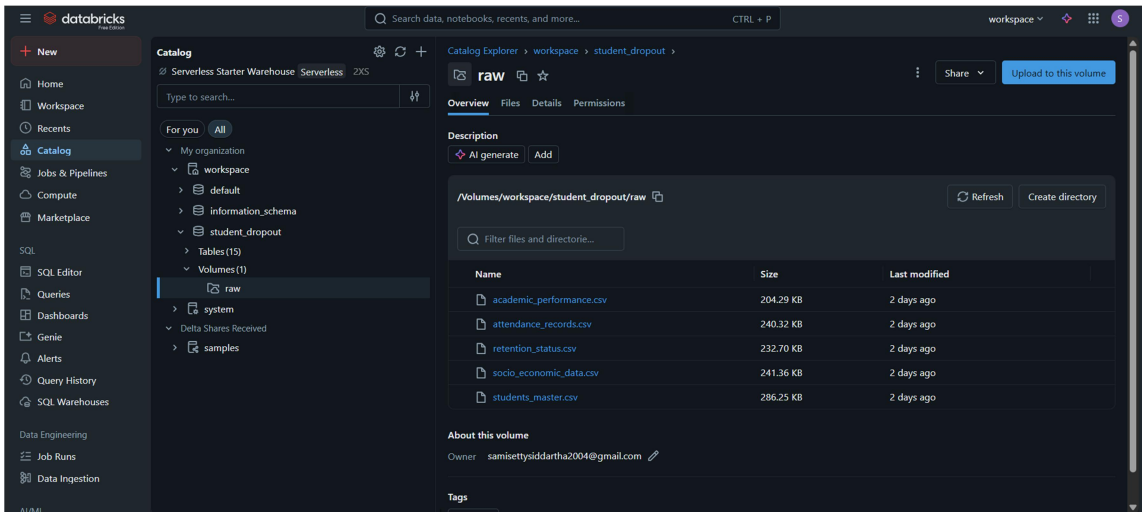
The system follows a **layered analytics architecture** with clear separation of responsibilities:

- **Databricks** – Data engineering and analytics
- **Apache Airflow** – Workflow orchestration
- **Power BI** – Visualization and decision support

## 6.2 Step-by-Step Data Flow

### Step 1: Raw Data Ingestion

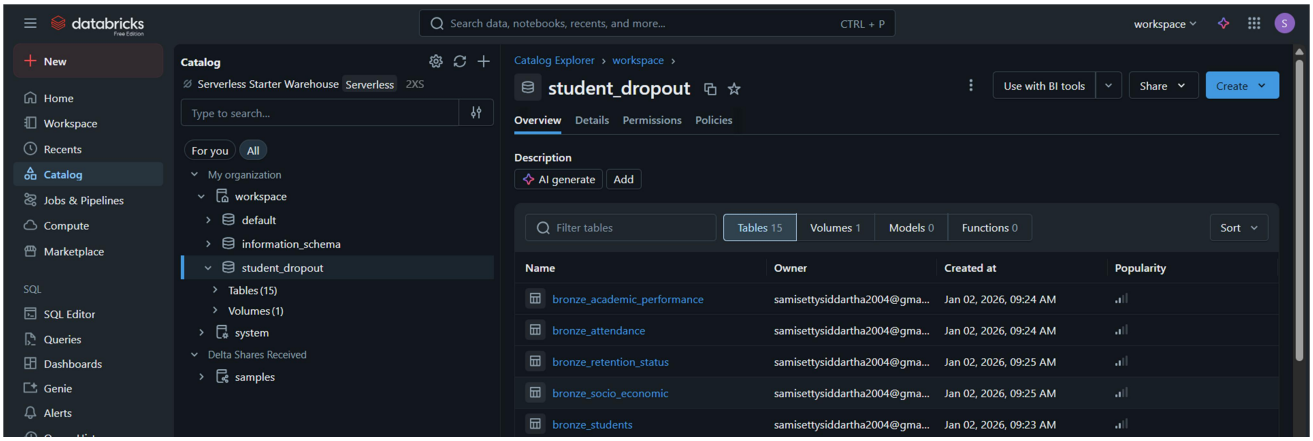
- CSV files are uploaded into Databricks
- Data is read using Spark
- No transformations are applied at this stage




### Step 2: Bronze Layer (Raw Storage)

- Raw datasets are stored as Delta tables
- Metadata columns added:
  - Ingestion timestamp
  - Source system
- Schema validation is performed

**Objective:** Preserve original data for traceability and auditing.



**Step 3: Silver Layer (Clean & Standardized Data)**

- Missing values handled
- Inconsistent categories standardized
- Derived metrics created:
    - Attendance percentage
    - Performance score
- Risk-related flags introduced

**Objective:** Prepare high-quality, analytics-ready data.

| | | |
|---|---|---|
| silver_academic_performance | samisettysiddartha2004@gma... | Jan 02, 2026, 09:26 AM |
| silver_attendance | samisettysiddartha2004@gma... | Jan 02, 2026, 09:26 AM |
| silver_retention_status | samisettysiddartha2004@gma... | Jan 02, 2026, 09:27 AM |
| silver_socio_economic | samisettysiddartha2004@gma... | Jan 02, 2026, 09:27 AM |
| silver_students | samisettysiddartha2004@gma... | Jan 02, 2026, 09:26 AM |

**Step 4: Gold Layer (Business Analytics)**

- All datasets joined using student_id
- Risk scores calculated
- Students classified into risk levels
- Intervention candidates identified

**Objective:** Produce business-ready analytics tables.

| | | |
|---|---|---|
| gold_attendance_vs_performance | samisettysiddartha2004@gma... | Jan 02, 2026, 09:28 AM |
| gold_dropout_trends | samisettysiddartha2004@gma... | Jan 02, 2026, 09:28 AM |
| gold_intervention_candidates | samisettysiddartha2004@gma... | Jan 02, 2026, 09:28 AM |
| gold_retention_overview | samisettysiddartha2004@gma... | Jan 02, 2026, 09:29 AM |
| gold_student_risk_profile | samisettysiddartha2004@gma... | Jan 02, 2026, 09:28 AM |

**Step 5: Workflow Orchestration**

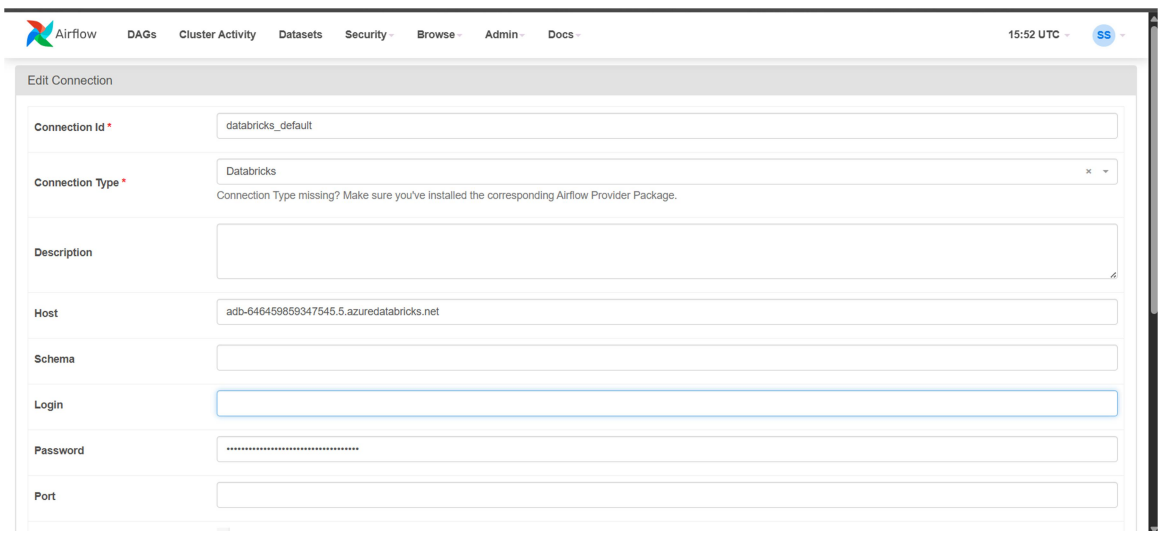
- Databricks Workflows manage execution order:

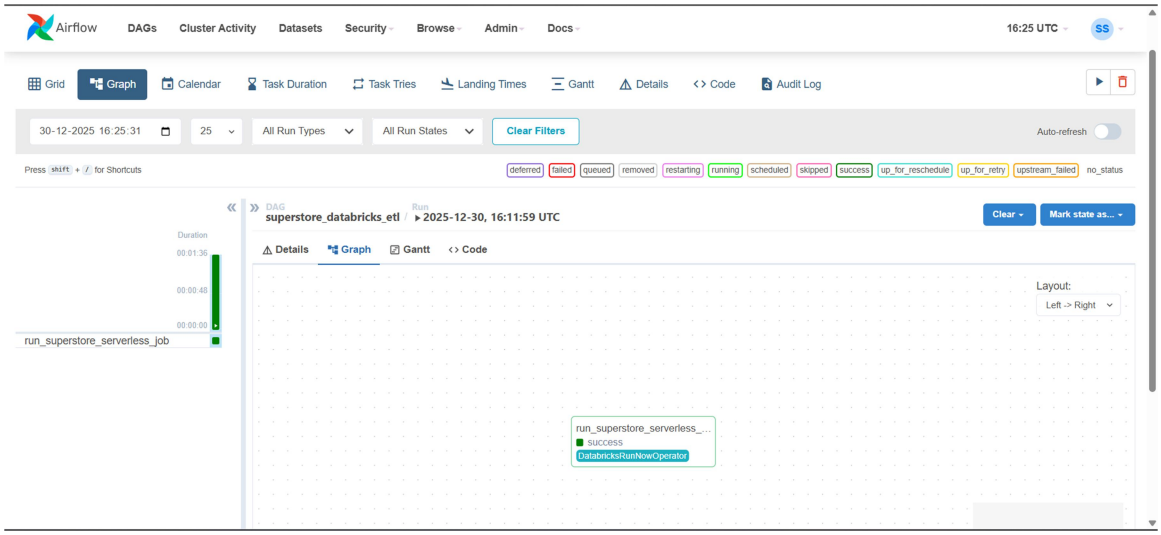
- Bronze → Silver → Gold → Validation
- Apache Airflow triggers and monitors the workflow

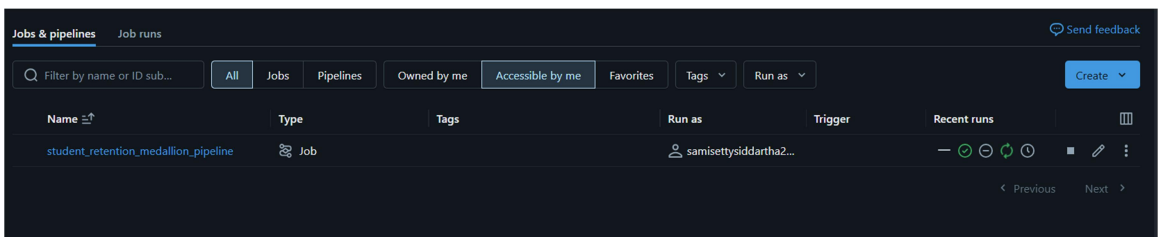
```
PS D:\capstone_project\Airflow> docker compose up -d
>>
[+] Running 3/3
 ✓ Container airflow-postgres-1          Running                                    0.0s
 ✓ Container airflow-airflow-scheduler-1 Started                                    0.7s
 ✓ Container airflow-airflow-webserver-1 Started                                    0.9s
PS D:\capstone_project\Airflow>
```

```
PS D:\capstone_project\Airflow> docker compose run --rm airflow-webserver airflow db init

/home/airflow/.local/lib/python3.8/site-packages/airflow/cli/commands/db_command.py:47 DeprecationWarning: `db init`
s deprecated.  Use `db migrate` instead to migrate the db and/or airflow connections create-default-connections to cr
ate the default connections
DB: postgresql+psycopg2://airflow:***@postgres/airflow
[2026-01-04T08:57:11.160+0000] {migration.py:216} INFO - Context impl PostgresqlImpl.
[2026-01-04T08:57:11.161+0000] {migration.py:219} INFO - Will assume transactional DDL.
INFO  [alembic.runtime.migration] Context impl PostgresqlImpl.
INFO  [alembic.runtime.migration] Will assume transactional DDL.
INFO  [alembic.runtime.migration] Running stamp_revision  -> 88344c1d9134
WARNI [airflow.models.crypto] empty cryptography key - values will not be stored encrypted.
Initialization done
```





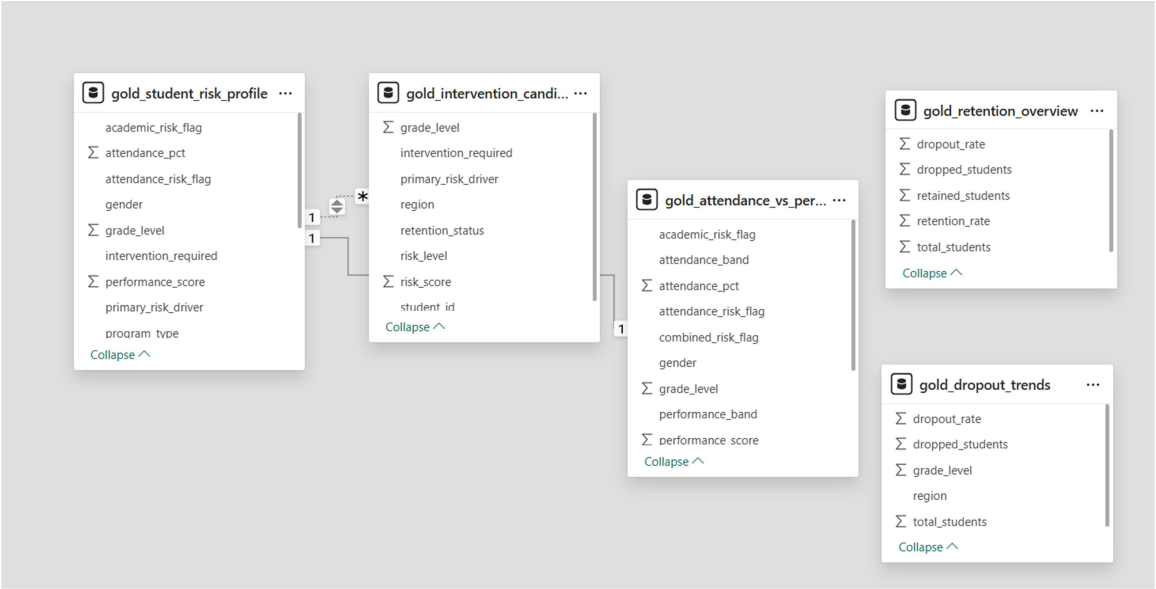




## Step 6: Visualization

- Power BI consumes Gold tables
- Dashboards provide insights and action points

## 7. RELATIONAL VIEW

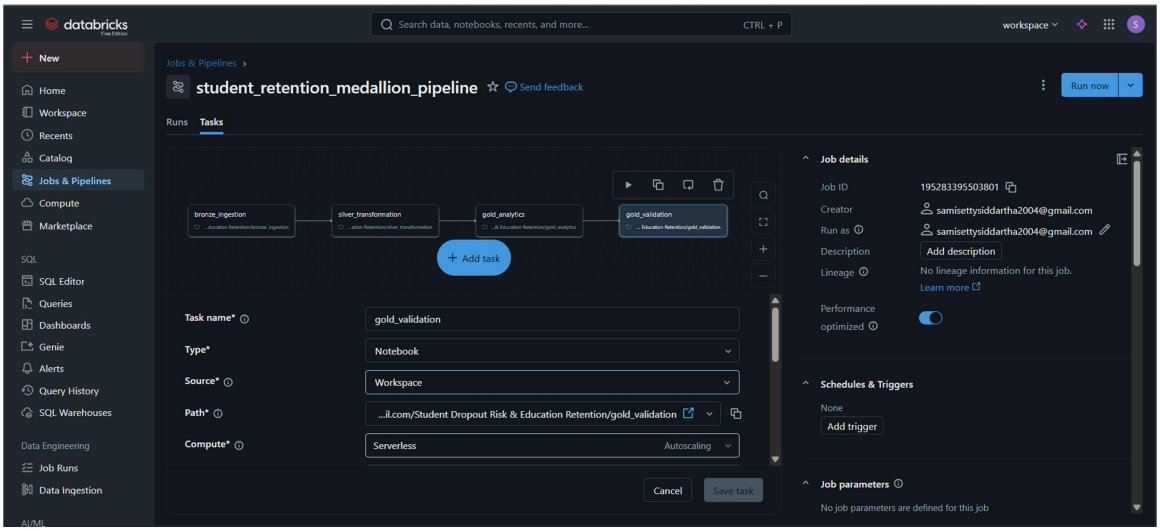
All datasets joined using student_id to form gold analytics tables.

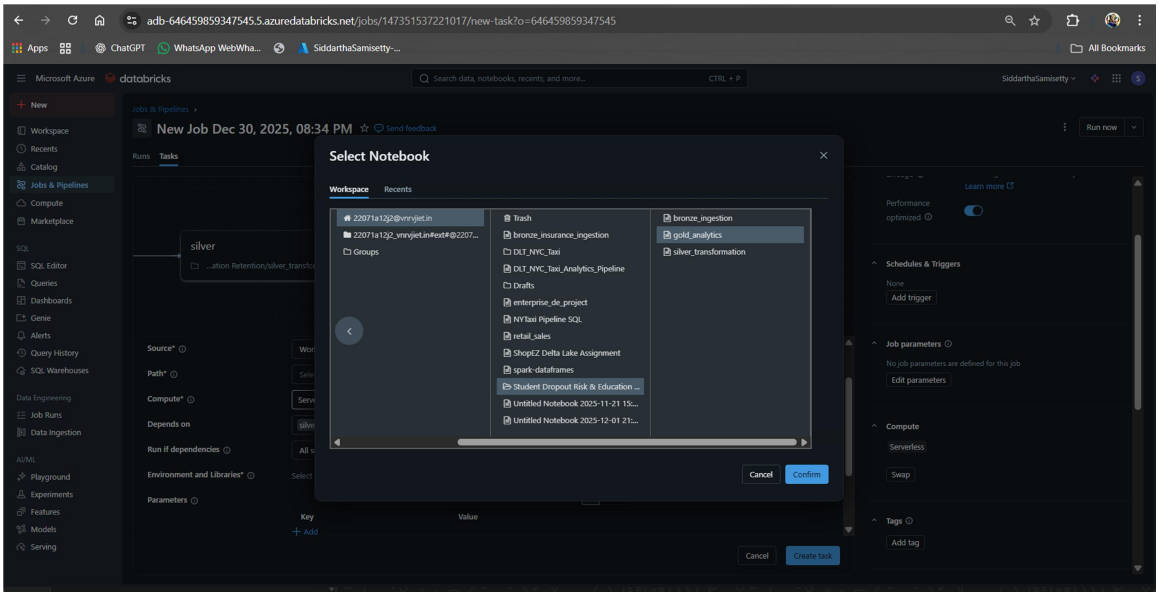



# 8. WORKFLOW ORCHESTRATION

## 8.1 Databricks Workflows

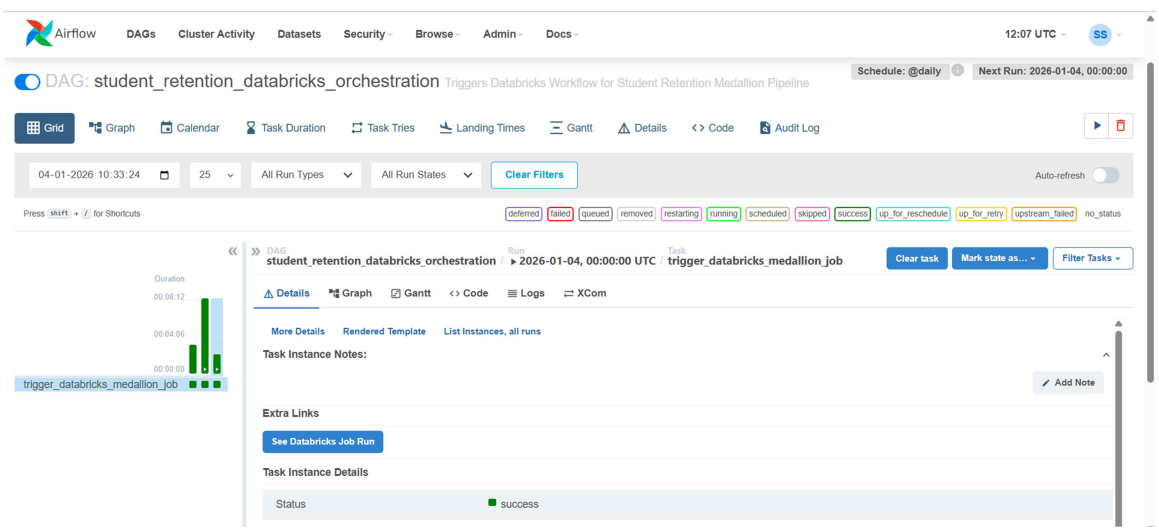
Bronze → Silver → Gold → Validation pipeline.






## 8.2 Apache Airflow

Triggers Databricks job, handles retries and monitoring.



# 9. Visualization & Reporting

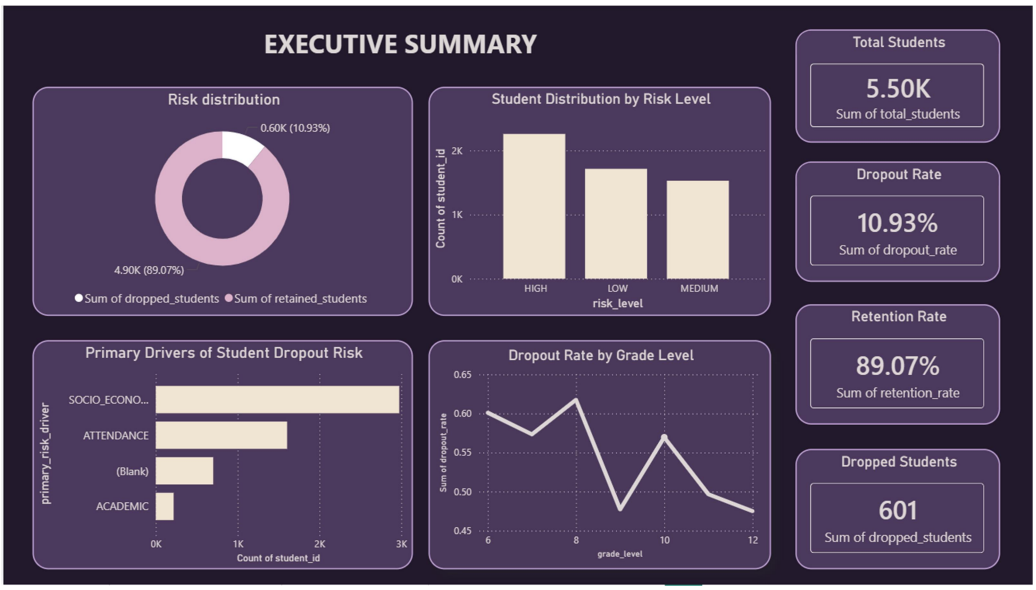
## 9.1 Executive Overview Dashboard

**Focus:**
High-level understanding of student retention health.

**Key Questions Answered:**

- What is the overall dropout rate?
- Which grades or regions are most affected?

**Why this page matters:**
Designed for **school leadership and administrators** to quickly assess institutional performance.



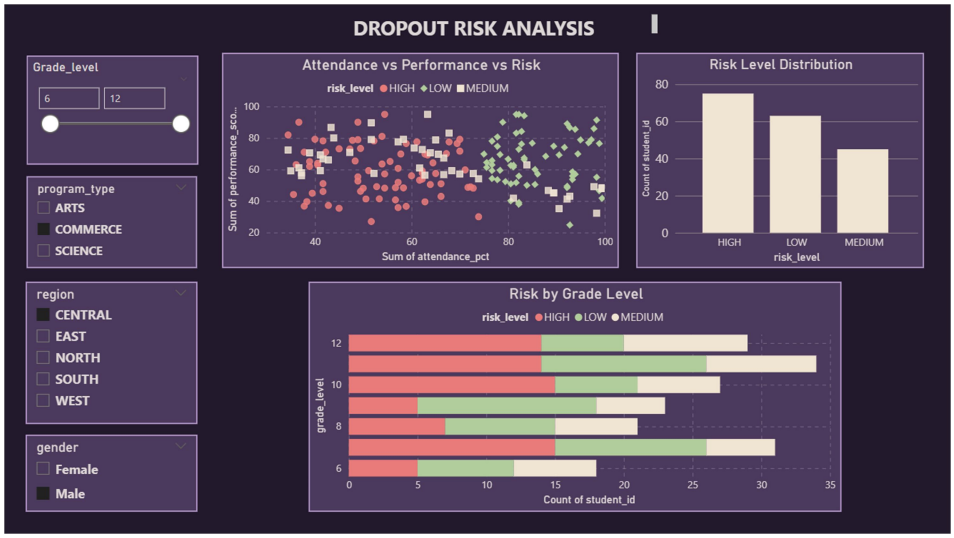

## 9.2 Dropout Risk Analysis Dashboard

**Focus:**
Understanding **who is at risk and why**.

**Key Insights:**

- Risk distribution across grades
- Attendance vs performance correlation
- Identification of high-risk clusters

**Significance:** Helps academic teams **diagnose root causes** of dropout risk.



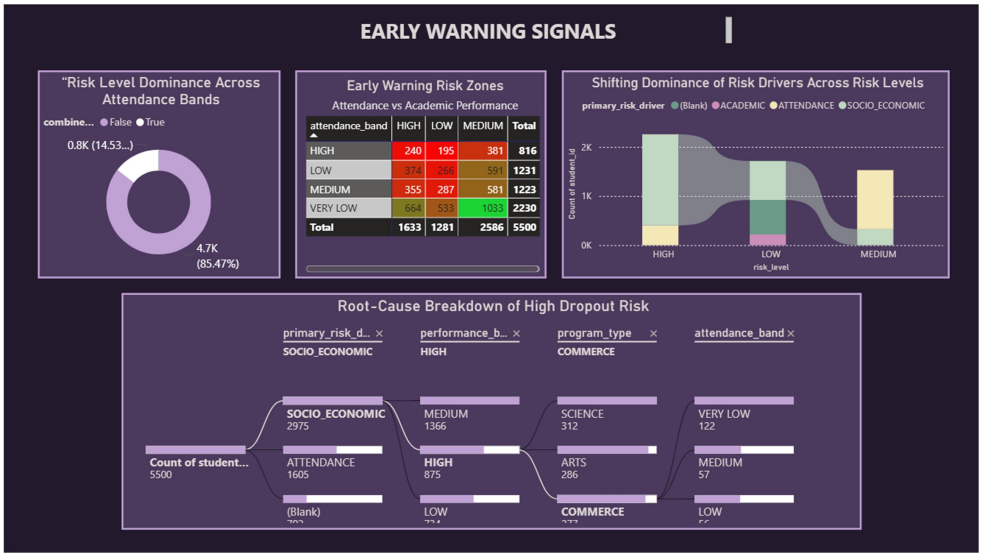

## 9.3 Early Warning Indicators Dashboard

**Focus:**
Detecting risk **before dropout occurs**.

**Key Indicators:**

- Attendance decline patterns
- Performance band shifts
- Combined risk dominance

**Why included:** Enables **preventive intervention**, not reactive response.



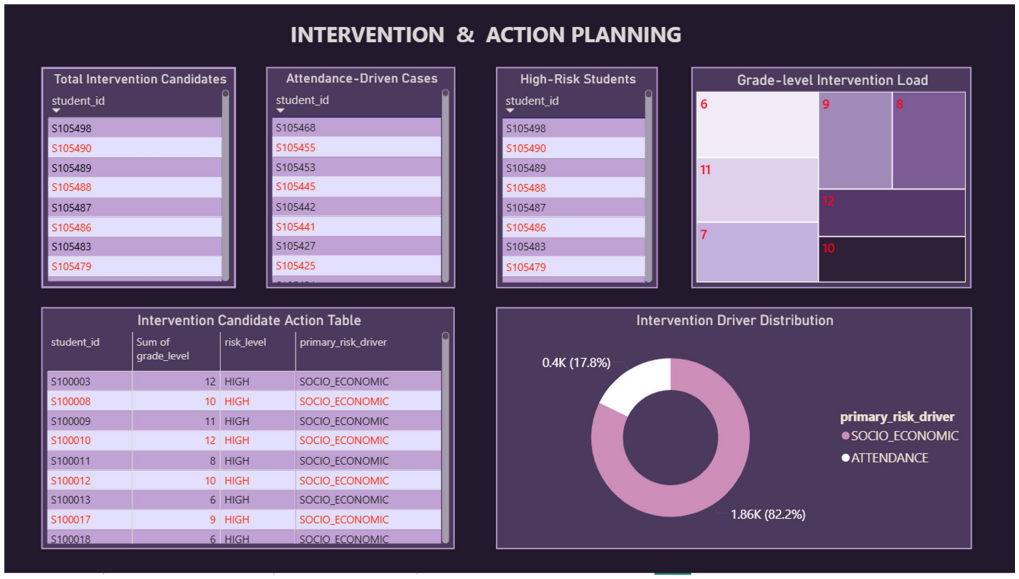

## 9.4 Intervention Planning Dashboard

**Focus:**
Turning insights into **actionable decisions**.

**Key Components:**

- Intervention KPI tables
- Grade-wise intervention load
- Student-level action list
- Primary risk driver distribution

**Significance:**
Supports counselors and educators in **resource allocation and targeted intervention planning**.



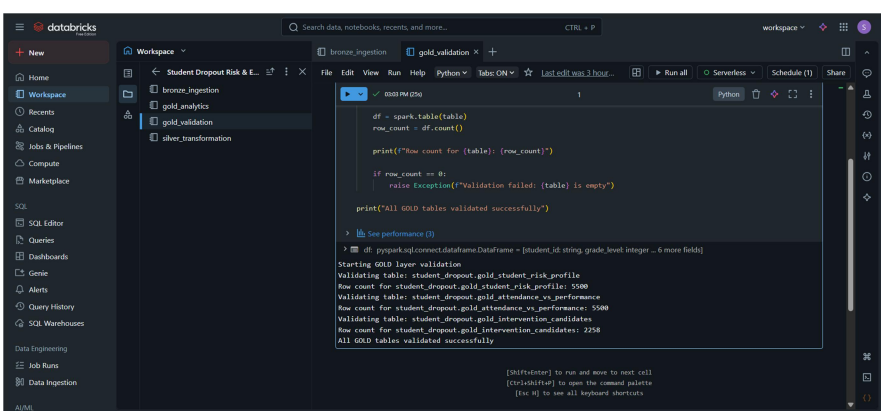

## 10. Testing & Validation

Testing and validation ensure data reliability and correctness.

Validation steps include:

- Ensuring Gold tables are not empty
- Verifying key identifiers are not null
- Validating risk flag logic
- Cross-checking dashboard values against Gold tables

A dedicated Gold validation notebook enforces these checks.



## 11. Machine Learning–Driven Dropout Risk Prediction

To enhance the analytical capabilities of the system beyond descriptive analytics, a **machine learning–based predictive layer** was introduced.

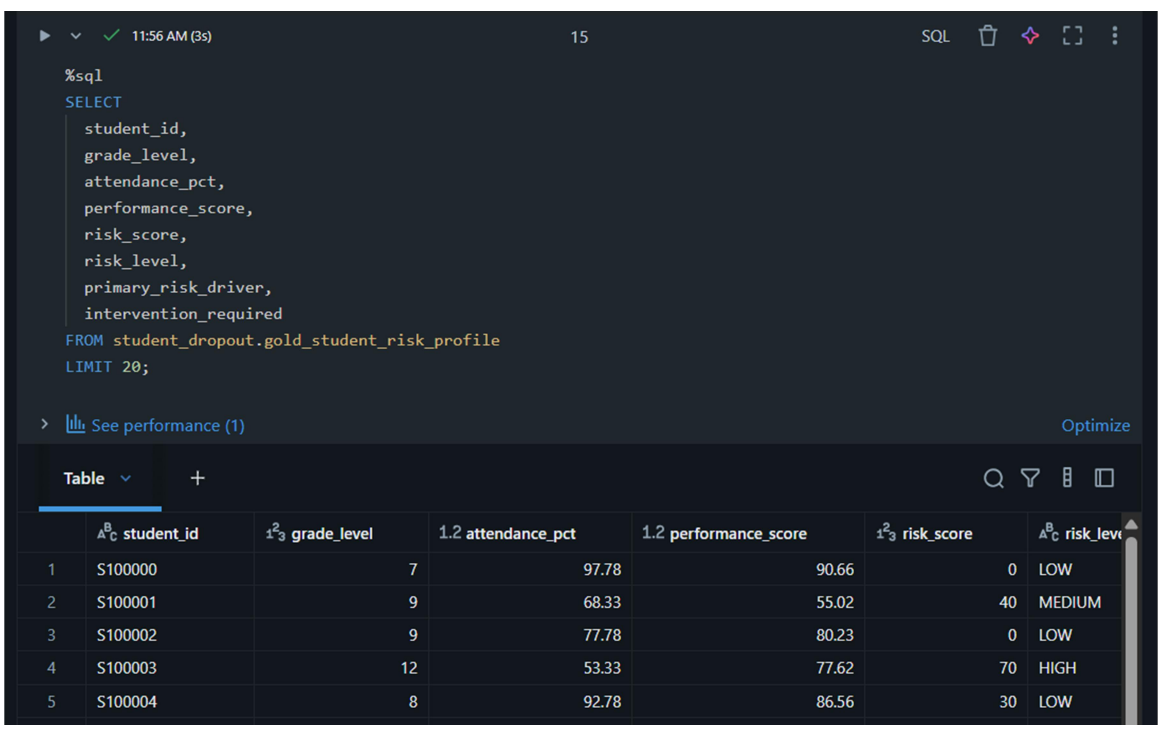
A **Logistic Regression model** was implemented using Spark MLlib to estimate the probability of student dropout. Logistic Regression was chosen due to its interpretability, suitability for binary classification problems, and acceptance in academic and enterprise environments.

### 11.1 Feature Selection

The model was trained using engineered features derived from the Gold layer, ensuring strict adherence to the Medallion Architecture principles. The selected features include:

- Attendance percentage
- Academic performance score
- Composite risk score
- Grade level
- Academic risk flag
- Attendance risk flag
- Socio-economic risk flag

The target variable used for prediction was **retention risk flag**, representing whether a student is at risk of dropping out.

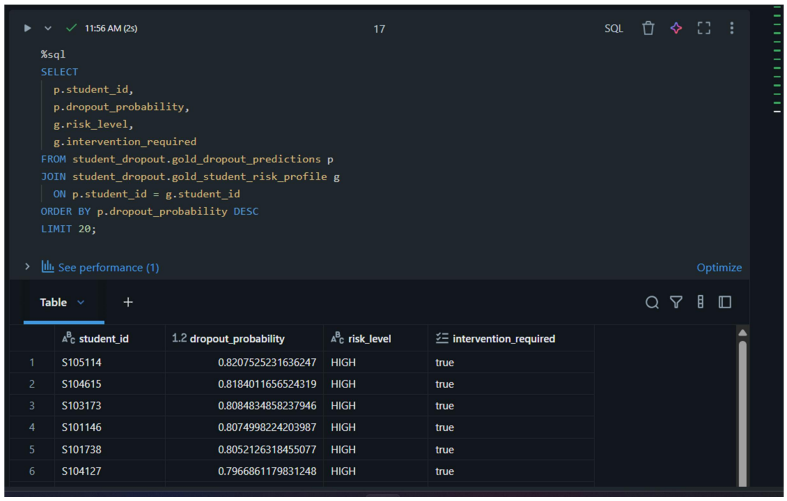



| | student_id | grade_level | attendance_pct | performance_score | risk_score | risk_level |
|---|---|---|---|---|---|---|
| 1 | S100000 | 7 | 97.78 | 90.66 | 0 | LOW |
| 2 | S100001 | 9 | 68.33 | 55.02 | 40 | MEDIUM |
| 3 | S100002 | 9 | 77.78 | 80.23 | 0 | LOW |
| 4 | S100003 | 12 | 53.33 | 77.62 | 70 | HIGH |
| 5 | S100004 | 8 | 92.78 | 86.56 | 30 | LOW |

### 11.2 Model Training and Evaluation

The dataset was split into training and testing sets using an 80:20 ratio. Model performance was evaluated using the **Area Under the ROC Curve (AUC)** metric to account for potential class imbalance.

The model achieved an **AUC score of approximately 0.90**, indicating excellent discriminatory power between at-risk and non-risk students. This confirms the effectiveness of the engineered Gold-layer features for predictive retention analytics.



```
%sql
SELECT
  p.student_id,
  p.dropout_probability,
  g.risk_level,
  g.intervention_required
FROM student_dropout.gold_dropout_predictions p
JOIN student_dropout.gold_student_risk_profile g
  ON p.student_id = g.student_id
ORDER BY p.dropout_probability DESC
LIMIT 20;
```

| | student_id | dropout_probability | risk_level | intervention_required |
|---|---|---|---|---|
| 1 | S105114 | 0.8207525231636247 | HIGH | true |
| 2 | S104615 | 0.8184011656524319 | HIGH | true |
| 3 | S103173 | 0.8084834858237946 | HIGH | true |
| 4 | S101146 | 0.8074998224203987 | HIGH | true |
| 5 | S101738 | 0.8052126318455077 | HIGH | true |
| 6 | S104127 | 0.7966861179831248 | HIGH | true |

```
%sql
SELECT
  student_id,
  dropout_probability
FROM student_dropout.gold_dropout_predictions
WHERE dropout_probability > 0.7
ORDER BY dropout_probability DESC;
```

| | student_id | dropout_probability |
|---|---|---|
| 1 | S105114 | 0.8207525231636247 |
| 2 | S104615 | 0.8184011656524319 |
| 3 | S103173 | 0.8084834858237946 |
| 4 | S101146 | 0.8074998224203987 |
| 5 | S101738 | 0.8052126318455077 |
| 6 | S104127 | 0.7966861179831248 |
| 7 | S102412 | 0.7953204965082985 |
| 8 | S105486 | 0.7942661167037893 |

```
11:53 AM (1s)                                    13
from pyspark.ml.evaluation import BinaryClassificationEvaluator

evaluator = BinaryClassificationEvaluator(
    labelCol="label",
    rawPredictionCol="rawPrediction",
    metricName="areaUnderROC"
)

auc = evaluator.evaluate(predictions)
print(f"AUC Score: {auc}")

AUC Score: 0.9049707602339183
```

## 12. Alerting Framework

To ensure timely action on identified risks, a **multi-layer alerting framework** was implemented. This framework distinguishes between **operational alerts** and **strategic governance alerts**, preventing alert fatigue while supporting different stakeholder needs.

## 12.1 Operational Alerts – ML-Based Email Notifications

Operational alerts are generated at the **student level** using ML-predicted dropout probabilities.

**Alert Logic**

An alert is triggered when a student's predicted dropout probability exceeds a defined threshold (e.g., $\geq 0.7$).

**Email Content**

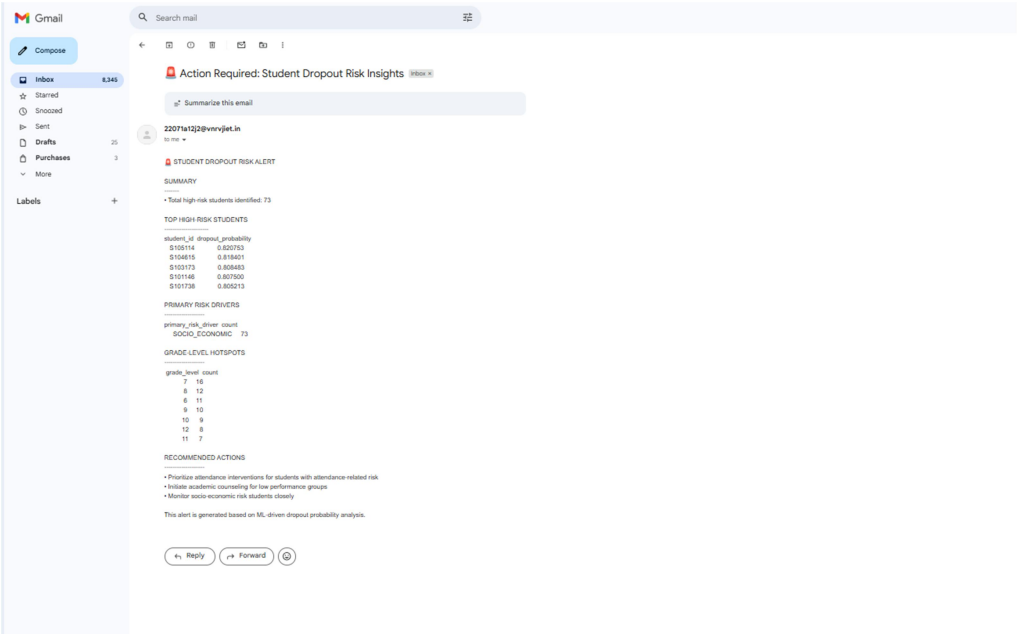
The alert email is designed to be **insight-rich and actionable**, containing:

- Total number of high-risk students detected
- List of top high-risk students
- Primary risk drivers (attendance, academic, socio-economic)
- Grade-level concentration of risk
- Recommended intervention focus areas

This ensures that counselors and academic teams receive **contextual information**, not just raw alerts.

**Security Considerations**

Due to workspace limitations, email credentials were injected securely at runtime within the notebook. In enterprise deployments, this mechanism can be replaced with Databricks Secret Scopes or centralized credential managers.




## 12.2 Strategic Alerts – Databricks SQL Governance Alerts

In addition to student-level alerts, a **Databricks SQL Alert** was implemented to monitor **institution-level risk trends**.

**Alert Objective**

The alert continuously monitors the **average dropout risk score** across the institution to detect **risk drift** over time.

**Baseline and Threshold Logic**

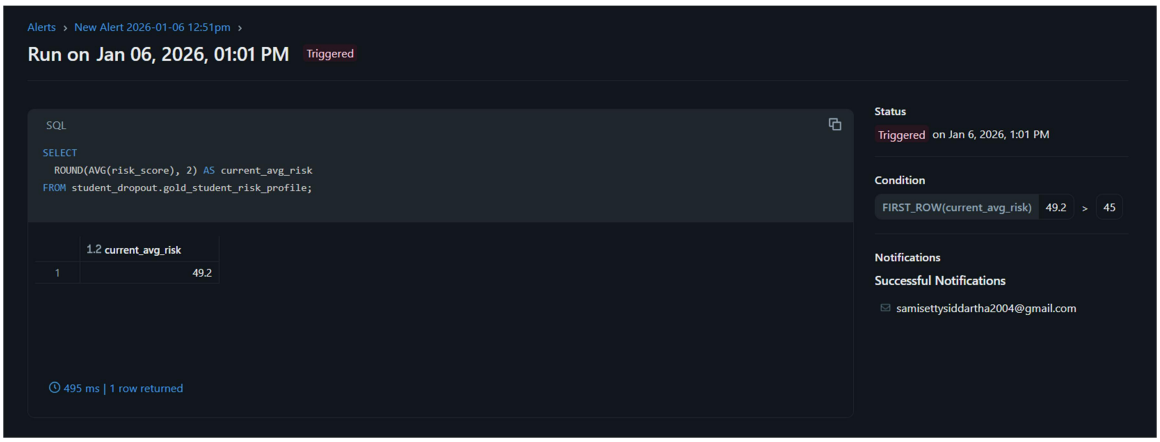
A baseline risk score was computed using historical Gold-layer data. The alert is triggered when the current average risk exceeds the baseline by a defined threshold, indicating systemic deterioration in student retention patterns.

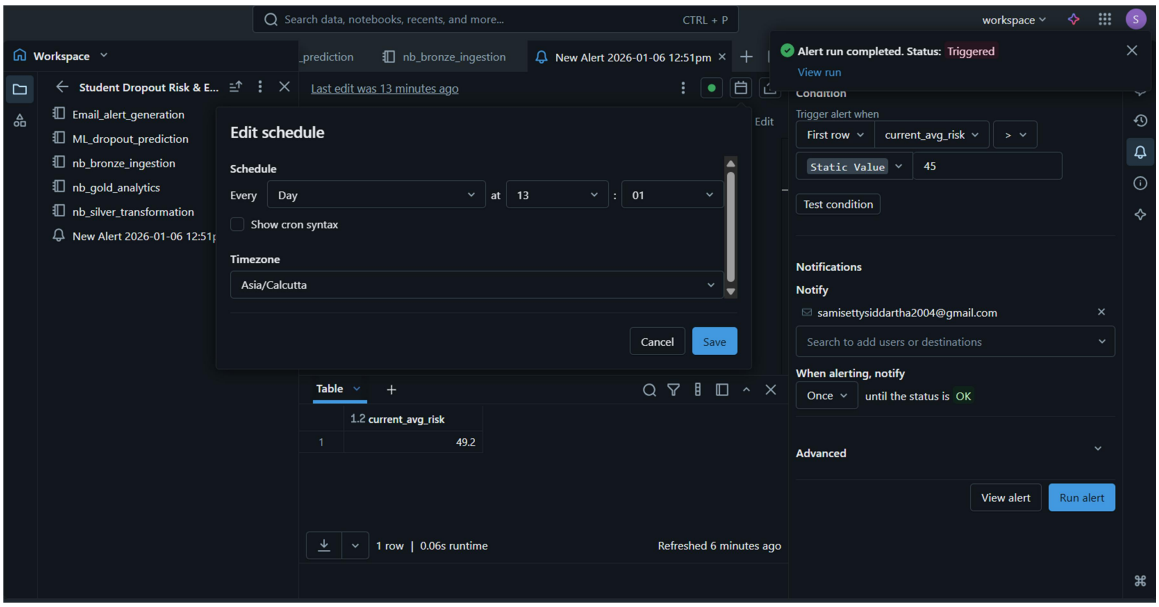
**Business Value**

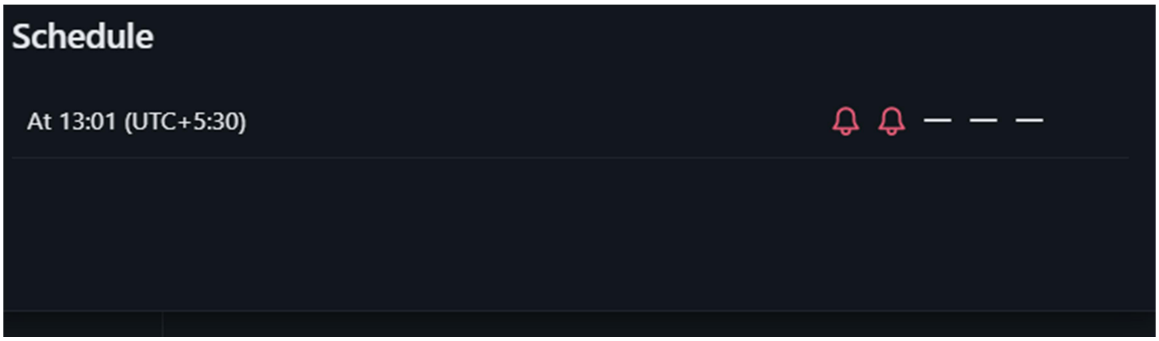
This alert supports:

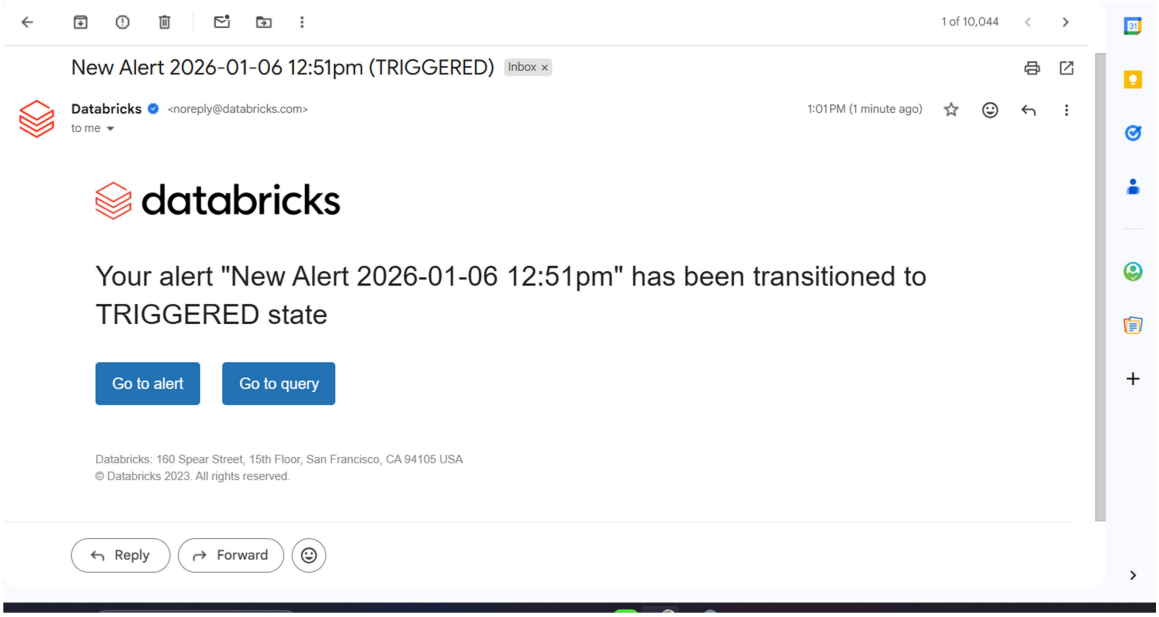
- Leadership-level decision-making
- Early identification of structural issues
- Proactive policy-level interventions

Unlike operational alerts, this governance alert focuses on **macro-level trends**, making it suitable for academic administrators and institutional planners.






## 13. Summary of Enhancements Beyond Core Requirements

In addition to the project's baseline requirements, the following advanced features were implemented:

- Machine learning–based early warning system
- Dual-layer alerting strategy (operational + strategic)
- Enterprise audit logging
- Risk drift monitoring using Databricks SQL Alerts
- ML-integrated dashboards and intervention insights

These enhancements significantly increase the project's real-world applicability and enterprise relevance.

## 14. Results & Business Impact

The system enables:

- **Early identification of at-risk students**
  The system highlights students who are likely to drop out at an early stage by analyzing attendance, academic performance, and socio-economic factors. This allows institutions to act before the situation becomes critical.
- **Focused and meaningful interventions**
  Each student is associated with a clear risk driver, such as attendance issues or academic struggles. This helps educators design targeted support strategies instead of applying generic solutions.

- **Better retention planning**
  Grade-wise and trend-based insights give administrators a clear understanding of where dropout risks are concentrated, enabling better planning of counseling resources and retention programs.
- **Reduced manual analysis and effort**
  Automated data pipelines, risk calculations, and alerts reduce dependency on manual data checks and spreadsheet-based analysis, saving time and minimizing human error.
- **Data-driven academic decision-making**
  Interactive dashboards and predictive insights allow educators and administrators to make informed decisions based on data rather than intuition.
- **Improved visibility at the institutional level**
  Governance-level alerts provide a high-level view of overall dropout risk trends, helping leadership monitor and respond to systemic issues proact

## 15. Challenges Faced & Solutions

| Challenge | Solution |
|---|---|
| Databricks cluster limitation | Used Databricks Workflows |
| Airflow CSRF errors | Configured local dev settings |
| Data inconsistencies | Handled in Silver layer |

## 16. Future Enhancements

While the current system delivers a complete, enterprise-ready analytics solution, several enhancements can further improve its scalability, intelligence, and institutional impact.

### 16.1 Real-Time Data Ingestion

The current batch-based ingestion can be extended to support **real-time or near-real-time streaming** using tools such as Kafka or cloud-native event streaming services. This would enable continuous monitoring of attendance and academic events, allowing interventions to occur even earlier.

### 16.2 Advanced Machine Learning Models

Future versions of the system may incorporate advanced models such as Random Forests, Gradient Boosting, or Neural Networks to capture complex non-linear relationships

between academic, behavioral, and socio-economic factors. Ensemble approaches could further improve predictive accuracy while maintaining explainability.

### 16.3 Automated Intervention Workflows

Currently, alerts notify stakeholders of potential risks. This can be enhanced by integrating **automated intervention workflows**, such as assigning counselors, scheduling meetings, or generating personalized academic support plans directly from alert triggers.

### 16.4 Feedback Loop for Model Improvement

A closed-loop feedback mechanism can be introduced where intervention outcomes are captured and fed back into the machine learning model. This would allow the model to continuously learn from real-world outcomes and improve its predictive performance over time.

### 16.5 Integration with Learning Management Systems (LMS)

Integrating the system with Learning Management Systems such as Moodle or Canvas would allow ingestion of behavioral data like assignment submissions, quiz activity, and engagement metrics, significantly enriching the feature set for dropout prediction.

### 16.6 Role-Based Dashboards and Access Control

Future enhancements can include role-based access control to ensure that different stakeholders (administrators, counselors, faculty) view dashboards tailored to their responsibilities, improving usability and data governance.

## 17. Conclusion

This project successfully delivers an **end-to-end Student Dropout Risk & Education Retention Analytics System** that integrates data engineering, analytics, machine learning, alerting, orchestration, and visualization into a unified solution.

By adopting an enterprise-grade Medallion Architecture, the system ensures scalability, data reliability, and maintainability. The transformation from raw data to analytics-ready Gold tables enables meaningful insights into academic performance, attendance behavior, and socio-economic factors influencing student retention.

The inclusion of **machine learning–based early warning mechanisms** elevates the system from descriptive analytics to predictive intelligence, allowing institutions to identify at-risk students before critical dropout events occur. The **dual-layer alerting**

**framework**, combining operational email alerts with strategic Databricks SQL alerts, ensures that both tactical and governance-level stakeholders receive timely and relevant information.

Additionally, enterprise features such as audit logging, workflow orchestration using Apache Airflow, and comprehensive validation checks enhance observability, accountability, and operational confidence.

Overall, this system demonstrates how modern data engineering and analytics practices can be applied to the education domain to shift from reactive analysis to **proactive, data-driven intervention strategies**. The project not only satisfies all specified requirements but also introduces advanced features that reflect real-world enterprise analytics implementations, making it both academically rigorous and practically re

.

.