CAPSTONE PROJECT

🎓 School Enrollment & Education Performance Intelligence Platform

By:

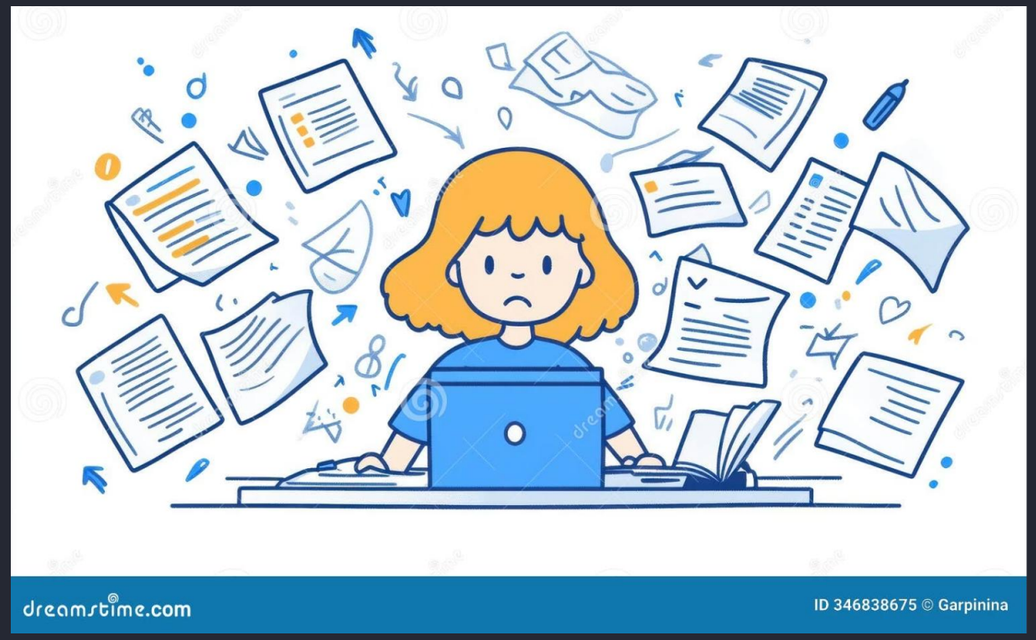
Naga Yaswanth Reddy Jonnala
VNRJIET

📊 Addressing Education Data Challenges

Education institutions often grapple with fragmented data, manual reporting, and a lack of real-time insights. These challenges directly impede effective decision-making and timely interventions, particularly concerning student enrollment and academic performance.

Current State Limitations:

- 📁 ***Scattered Data:*** *Enrollment and performance data are dispersed across numerous CSV files, leading to labor-intensive manual reporting that consumes 2-3 days per month.*
- 📝 ***Manual Reporting:*** *Current processes are time-consuming and limited to basic Excel reports, preventing in-depth trend analysis and predictive insights.*
- ❌ ***No Centralized Analytics:*** *The absence of a unified analytics system results in decision-making delays of 1-2 weeks, hindering responsiveness.*
- ⚠️ ***Identification Difficulty:*** *Significant challenges exist in identifying critical enrollment trends and early indicators of dropout risks, contributing to an average dropout rate of 31.74% that requires proactive intervention strategies.*



Business Impact:

- *Delayed interventions for at-risk students*
- *Missed opportunities for capacity planning and resource allocation*
- *Reactive instead of proactive decision-making in critical areas*

> ***Critical Need:*** *An automated, scalable analytics platform is essential to transform data into actionable intelligence and drive better educational outcomes.*

🎯 Project Objective: Building a Production-Grade Analytics Pipeline

This project aims to establish a robust, automated, and scalable data pipeline that transforms raw education data into actionable insights, enabling data-driven decision-making for school administrators and stakeholders.

1.Ingest Raw Data

Automate the ingestion of raw education data from diverse sources, including 7,000+ enrollment and 7,000+ performance records from CSV files.

2.Clean & Validate Datasets

Implement rigorous data cleaning and validation processes to remove duplicates, handle null values, and filter outliers, ensuring data integrity.

3.Generate KPI-Ready Analytics

Develop logic to generate key performance indicator (KPI)-ready analytics, including enrollment trends, precise dropout rates, and comprehensive academic performance metrics.

4.Automate Execution

Orchestrate the entire pipeline using Apache Airflow, scheduling daily runs and integrating email notifications for operational transparency.

5.Provide Interactive Dashboards

Create intuitive and interactive Power BI dashboards that offer real-time insights and drill-down capabilities for decision-makers.

Success Metrics:

- ***Pipeline Success Rate:*** *Achieve a 100% success rate for all automated pipeline runs.*
- ***Data Quality Score:*** *Maintain a data quality score of 98.5% or higher post-cleaning and validation.*
- ***Execution Time:*** *Ensure the pipeline completes its entire execution within approximately 5 minutes.*

⚙️ End-to-End Data Flow: From Raw Data to Insights

This comprehensive solution follows an industry-standard Medallion Architecture, ensuring data quality, consistency, and optimized access for analytics. Apache Airflow orchestrates the entire process, delivering a fully automated, scalable, and repeatable pipeline.

Raw CSV Files

Initial data sources comprising 3 datasets and over 7,000 records:

- `school_master.csv`
- `enrollment.csv`
- `performance.csv`

🟫 BRONZE LAYER: Raw Ingestion

Ingests raw data with schema enforcement and identifier validation. Outputs `bronze_*` *Delta tables, preserving original data structure.*

🟣 SILVER LAYER: Cleaned & Validated

Applies deduplication using window functions, outlier removal (0-100 range for scores), and standardization (Gender, Region). Outputs `silver_*` *Delta tables, creating a trusted data source.*

🟡 GOLD LAYER: Analytics Ready

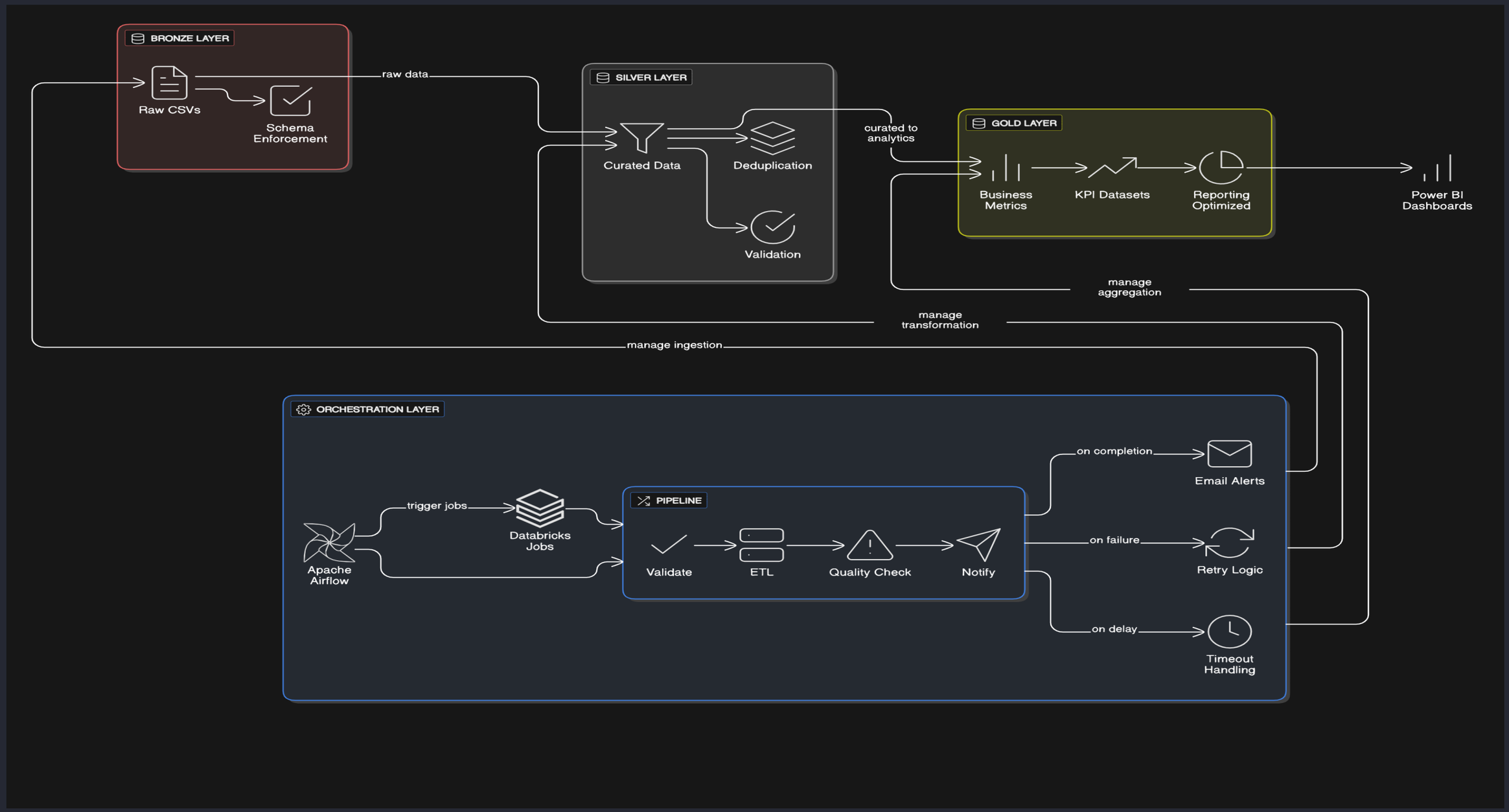
Aggregates business metrics and calculates KPIs. Outputs 2,955 enrollment and 720 performance analytics records, optimized for reporting.

📊 POWER BI DASHBOARDS

Two interactive dashboards with dynamic filters provide real-time insights for decision-makers.

Orchestration: *The entire process is orchestrated daily at 2 AM by Apache Airflow, ensuring full automation and reliability.*

Architecture



Comprehensive Technology Stack

A carefully selected suite of modern data technologies underpins this platform, each chosen for its specific strengths in handling large-scale data processing, orchestration, and visualization.

Processing	PySpark (Apache Spark)	Distributed transformations for large datasets, ensuring efficiency and scalability.
Platform	Databricks	A unified, serverless analytics platform for collaborative data science and engineering.
Orchestration	Apache Airflow 2.8	Manages and automates complex data workflows, scheduling daily runs and handling dependencies.
Visualization	Power BI	Creates interactive dashboards and reports for business users to gain insights.
Containerization	Docker	Ensures reproducible environments for development and deployment, simplifying setup.
Language	Python 3.9+	The primary development language for scripting, data manipulation, and custom logic.

Why These Technologies?

- **Spark:** Crucial for handling large-scale data (259K+ records) with high performance.
- **Airflow:** Provides robust, production-grade workflow automation and monitoring.
- **Power BI:** Offers user-friendly and powerful business intelligence capabilities.
- **Docker:** Facilitates consistent deployments across different environments.

📁 Raw Datasets: The Foundation of Our Analytics

The project leverages three core CSV datasets stored in Databricks Volumes, providing a rich source of information on school administration, student enrollment, and academic performance. These datasets are the starting point for our ETL pipeline.

1. School Master

(120 records)

- ***Content:*** *Contains essential metadata for each school, including region, district, student capacity, and school type.*
- ***Role:*** *Serves as a crucial reference table for joining with enrollment and performance data, providing contextual information for analysis.*

2. Student Enrollment

(7,000 records)

- ***Content:*** *Detailed records of student enrollment categorized by academic year, grade level, and gender.*
- ***Coverage:*** *Includes 6 years of historical data, from 2019 to 2024, enabling trend analysis.*

3. Student Performance

(7,000 records)

- ***Content:*** *Academic scores, attendance percentages, and flags indicating dropout risk for individual students.*
- ***Linkage:*** *Directly linked to schools and enrollment records to provide a holistic view of student progress.*

Common Data Issues Encountered:

The raw datasets presented several typical data quality challenges that required robust cleaning and validation:

- ***Null Values:*** *Critical fields often contained missing values.*
- ***Inconsistent Labels:*** *Varied representations for the same categorical data (e.g., "M/F" vs "Male/Female").*
- ***Missing/Invalid Grade Levels:*** *Entries like "grade_level = 0" indicating unknown or erroneous data.*
- ***Outliers:*** *Academic scores or attendance percentages outside the logical 0-100 range.*
- ***Duplicate Records:*** *Redundant entries that could skew analytical results.*

🟫 Bronze Layer: Raw Data Ingestion

The Bronze layer serves as the initial landing zone for raw data. Its primary purpose is to ingest data with minimal transformation, preserving the original state while enforcing schema for structural integrity. This layer acts as an immutable, historical record of the source data.

Key Actions Performed:

- ***Schema Enforcement:*** *Explicitly defines and enforces data types for all incoming columns, preventing schema inference issues.*
- ***Record Count Validation:*** *Verifies the expected number of records (e.g., 7,000 per dataset) to ensure complete ingestion.*
- ***Identifier Validation:*** *Ensures critical identifiers, such as* `school_id`*, are not null, maintaining fundamental data linkages.*
- ***Referential Integrity Checks:*** *Basic checks to identify potential orphan records or mismatches early in the pipeline.*

Code Example: Ingesting Enrollment Data

```
df_enrollment = spark.read \
  .format("csv") \
  .option("header", "true") \
  .schema(enrollment_schema) \
  .load(ENROLLMENT_PATH)

df_enrollment.write \
  .format("delta") \
  .mode("overwrite") \
  .saveAsTable("bronze_student_enrollment")
```

What We DON'T Do Here:

- ***No Data Transformations:*** *Data is loaded as-is without any modification of values.*
- ***No Joins or Aggregations:*** *Complex data manipulations are deferred to subsequent layers.*
- ***No Cleaning or Filtering:*** *Data quality issues are not addressed at this stage; raw values are retained.*

Silver Layer: Cleaning, Validation, and Curation

The Silver layer is where raw data is transformed into a clean, standardized, and validated dataset. This layer is crucial for building trust in the data, making it ready for reliable analytics and ensuring all downstream processes operate on high-quality information.

Transformations Applied:

1. ***Null Handling:*** *Records with missing critical identifiers (*`school_id` *or* `year`*) are dropped to maintain data integrity.*
2. ***Standardization:*** *Inconsistent labels are harmonized (e.g., "M" to "Male", "F" to "Female", "NORTH" to "North").*
3. ***Deduplication:*** *Utilizes Spark's window functions (*`partitionBy school+year+grade+gender`*) to identify and remove duplicate records, ensuring uniqueness.*
4. ***Outlier Filtering:*** *Academic scores and attendance percentages are filtered to ensure they fall within the logical 0-100 range.*
5. ***Grade-Level Filtering:*** *Invalid grade levels (e.g.,* `grade_level = 0`*) are removed.*
6. ***Referential Integrity:*** *Ensures all enrollment records correctly link to valid schools, preventing orphaned data.*

Code Example: Deduplication

```
window_spec = Window.partitionBy(
  "school_id", "academic_year",
  "grade_level", "gender"
).orderBy("timestamp_column") # Add an order to pick specific
duplicate

df_dedup = df.withColumn("row_num",
row_number().over(window_spec)) \
  .filter(col("row_num") == 1)
```

Quality Check Results:

- ***Data Completeness:*** *Achieved 98.5% completeness after cleaning.*
- ***No Duplicates:*** *All identified duplicates were successfully removed.*
- ***Valid Ranges:*** *All numerical values are now within their valid ranges.*

🟡 Gold Layer: Business-Ready Analytics & KPIs

The Gold layer is the final stage of the data pipeline, dedicated to creating business-ready metrics and key performance indicators (KPIs). This layer aggregates and transforms the clean data from the Silver layer into highly consumable formats optimized for direct use by business intelligence tools and decision-makers.

Business Metrics Calculated:

- ***Total Enrollment:*** *Calculated at 259,000 students across all schools and years.*
- ***Year-over-Year Growth:*** *Utilized* `lag()` *window functions to compute annual enrollment and performance growth rates.*
- ***Average Academic Score:*** *Determined to be 67.37 across all student performance records.*
- ***Average Attendance:*** *Calculated at 80.16% across the dataset.*
- ***Dropout Rate:*** *Established at 31.74% (students at-risk / total students), a critical KPI for intervention.*
- ***High-Risk Schools:*** *Identified schools exhibiting a dropout rate exceeding 20%.*
- ***Regional Analysis:*** *Provides aggregated enrollment and performance insights segmented by geographical region.*

Output Tables:

The Gold layer generates two primary analytics tables:

- `enrollment_analytics`*: Contains 2,955 records with aggregated enrollment trends and insights.*
- `performance_analytics`*: Comprises 720 records detailing academic performance and dropout metrics.*

Code Example: Performance Metrics Aggregation

```
gold_performance_metrics = df_performance \
  .groupBy("school_id", "academic_year", "grade_level") \
  .agg(
    avg("average_score").alias("avg_score"),
    avg("attendance_percentage").alias("avg_attendance"),
    sum("dropout_risk_flag").alias("students_at_risk")
  ) \
  .withColumn("total_students", sum(lit(1)).over(window_spec)) \
  .withColumn("dropout_risk_percent",
    (col("students_at_risk") / col("total_students")) * 100
  )
```

Apache Airflow Orchestration: Powering the Education ETL Pipeline

Orchestrating complex data pipelines is crucial for timely and accurate insights. Our Education ETL Production DAG (Directed Acyclic Graph) in Apache Airflow ensures seamless, automated data flow from source validation to final alerts. This robust orchestration layer acts as the control center, managing dependencies, monitoring progress, and handling potential failures with resilience.

1

Validate Sources

Initiates the pipeline by checking the integrity and availability of raw data sources. This critical first step prevents downstream failures from malformed or missing input files.

2

Databricks ETL Job

Triggers the core ETL process on Databricks, transforming raw data (Bronze) into clean, structured layers (Silver and Gold) using Spark's distributed computing power.

3

Quality Checks

Executes comprehensive data quality validations on the transformed data, ensuring accuracy, completeness, and consistency before it's consumed by reporting tools.

4

Email Alert

Sends automated notifications to stakeholders, confirming pipeline completion or alerting them to any issues, maintaining transparency and accountability.

Beyond the sequential flow, the DAG incorporates production-grade features to ensure reliability and maintainability. These features are fundamental for any enterprise-level data operation.

Retries

Configured for 2 attempts with a 5-minute delay, mitigating transient failures and ensuring pipeline completion without manual intervention.

Timeout

A 2-hour safety limit on the entire DAG prevents runaway processes, conserving compute resources and signaling potential issues promptly.

Logging

Detailed, task-level execution logs provide granular visibility into each step, aiding in debugging and performance analysis.

Monitoring

Real-time status tracking via Airflow's UI allows operators to oversee pipeline health and identify bottlenecks immediately.

Scheduling

The pipeline is scheduled to run daily at 2:00 AM UTC, ensuring that stakeholders receive fresh data at the start of each business day.

Notifications

Automated email notifications provide instant feedback on pipeline success or failure, reducing the need for constant manual checks.

> 📧 ***Email Notification**: "✅ Education ETL Completed! Pipeline completed successfully! Education ETL job ran without issues."*

Multi-Layer Observability: Comprehensive Monitoring & Logging Strategy

1 Airflow Task-Level Logs

Each task within the Airflow DAG generates detailed logs, offering a granular view of its execution. This allows for pinpointing exact failure points and understanding the behavior of individual components:

- ***Validate task:*** *Records outcomes of file existence checks and data schema validations.*
- ***ETL task:*** *Logs the status of Databricks job triggers, including start and end times.*
- ***Quality task:*** *Captures the results of all validation rules applied to the transformed data.*
- ***Notify task:*** *Confirms successful email delivery or logs any notification failures.*

2 Databricks Job Logs

The core ETL processes running on Databricks provide their own rich set of logs, essential for understanding Spark job performance and data transformation specifics:

- ***Spark execution logs:*** *Detailed logs from Spark drivers and executors.*
- ***Bronze/Silver/Gold layer metrics:*** *Track data volume and schema changes across layers.*
- ***Record counts at each stage:*** *Essential for verifying data consistency and detecting data loss.*
- ***Error traces:*** *Provide stack traces for job failures, accelerating root cause analysis.*

3 Post-Run Data Quality Checks

After the ETL process, automated data quality checks are performed to ensure the reliability and usability of the final data products. These checks are crucial for maintaining data integrity:

- ***Row count validation:*** *Compares expected vs. actual row counts to detect data discrepancies.*
- ***Null value checks:*** *Identifies critical columns with an unacceptable percentage of nulls.*
- ***Range validation:*** *Ensures numerical data falls within predefined acceptable ranges (e.g., grades 0-100).*
- ***Referential integrity:*** *Verifies relationships between tables are maintained correctly.*

4 Audit Metrics Tables

For long-term analysis and operational insights, we maintain dedicated audit tables. These tables track metadata about pipeline runs and data quality over time:

- ***Pipeline execution history:*** *Records start/end times, status, and duration for every run.*
- ***Duration tracking:*** *Helps identify performance degradation or improvements over time.*
- ***Data quality scores:*** *Aggregated metrics on data quality, allowing for trend analysis and proactive issue detection.*

Power BI Dashboards: Interactive Analytics for Education Decision-Makers

Power BI dashboards transform raw education data into actionable insights, providing key stakeholders with interactive tools to monitor trends, assess performance, and make informed decisions. These dashboards are designed for clarity, interactivity, and real-time data access.

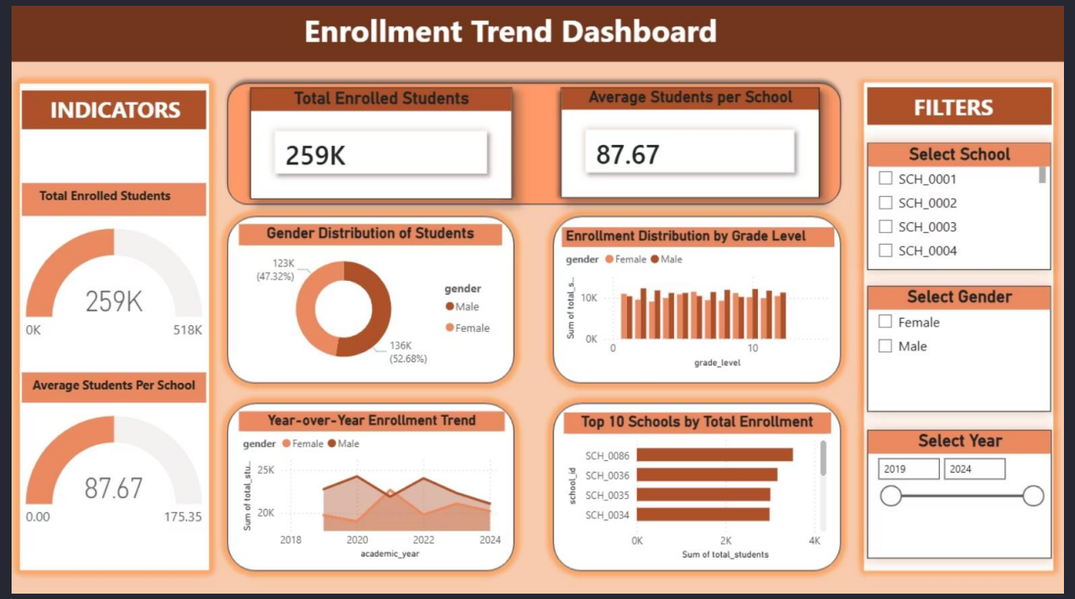
Dashboard 1: Enrollment Trends

This dashboard provides a comprehensive overview of student enrollment dynamics, crucial for resource allocation and strategic planning.

- ***Total enrollment:*** *259,000 students currently enrolled.*
- ***Avg students per school:*** *An average of 87.67 students per institution provides context on school size.*
- ***Gender split:*** *A detailed breakdown of 52.68% Female / 47.32% Male enrollment figures.*

Visualizations:

- *Year-over-year enrollment trend (2019-2024) allows for historical analysis.*
- *Gender distribution (donut chart) offers a clear visual breakdown.*
- *Grade-wise enrollment (bar chart) highlights concentrations across different academic levels.*
- *Top 10 schools by enrollment showcases institutions with the largest student bodies.*



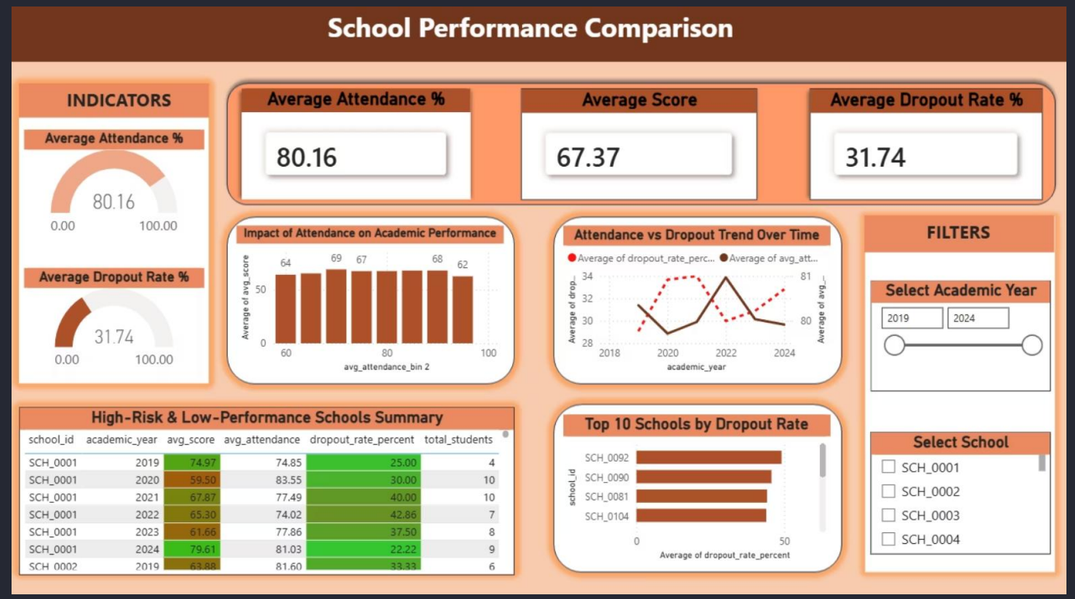
Dashboard 2: School Performance

Focused on critical academic and attendance metrics, this dashboard helps identify areas of strength and concern within the school system.

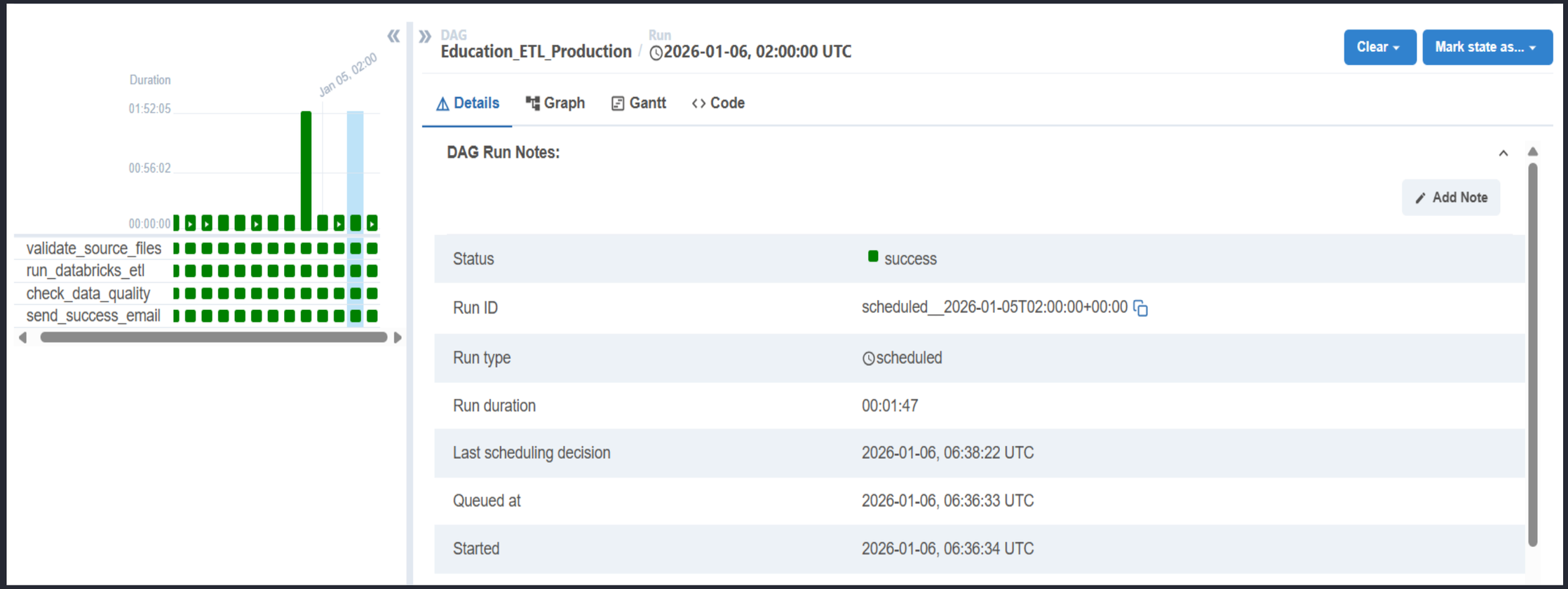
- ***Avg academic score:*** *System-wide average of 67.37.*
- ***Avg attendance:*** *Average attendance rate of 80.16%.*
- ***Avg dropout rate:*** *A significant average dropout rate of 31.74%, signaling areas for intervention.*

Visualizations:

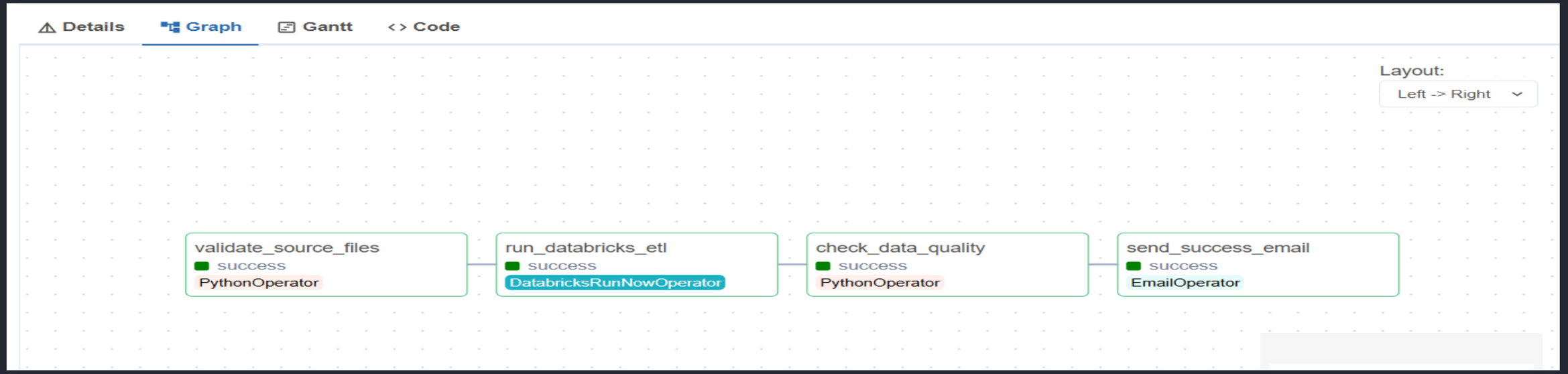
- *Attendance impact on performance (correlation chart) reveals key relationships.*
- *Dropout vs attendance trend over time provides insights into longitudinal patterns.*
- *Top 10 high-risk schools (>30% dropout) allows for targeted support.*
- *Heat map with conditional formatting quickly highlights performance variations across regions or schools.*

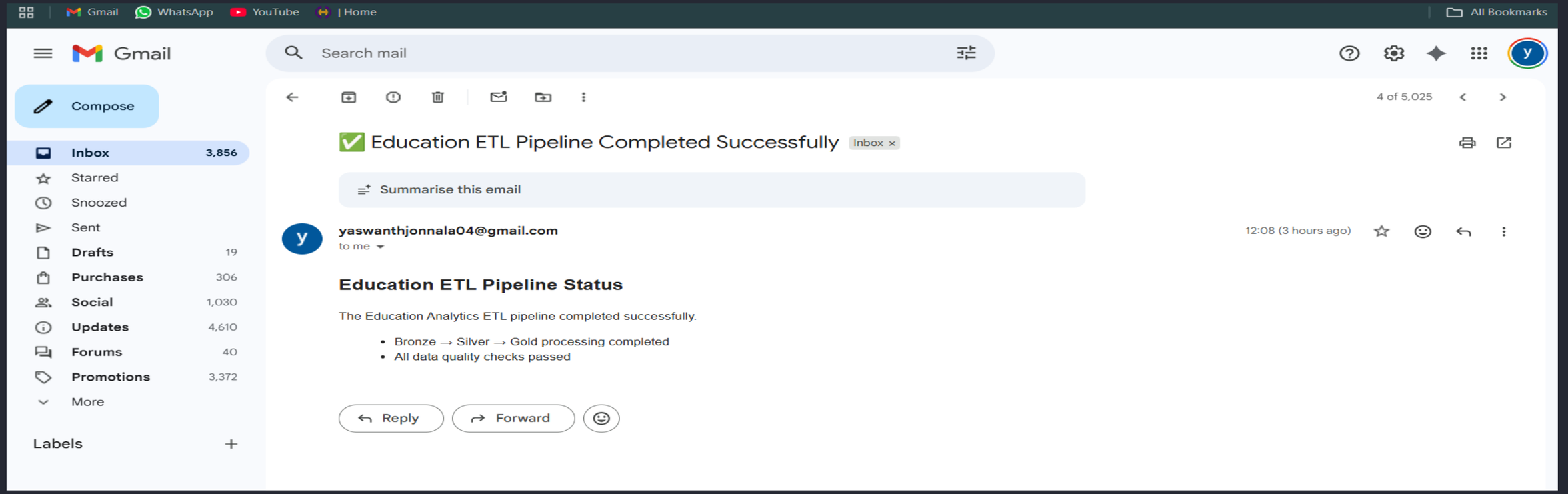


Workflow Automation



Workflow Automation





Key Outcomes & Learnings: Delivering a Robust Education Analytics Platform

This project has not only delivered a functional data platform but has also yielded significant quantitative results and qualitative impacts, demonstrating the power of modern data engineering to transform educational decision-making. The journey also provided invaluable learnings, solidifying our expertise in building production-grade solutions.

Project Outcomes

Quantitative Results:

- *Successfully processed **259,000+ student records** from 120 diverse schools, centralizing fragmented data.*
- *Generated **2,955 detailed enrollment analytics records**, providing a granular view of student demographics and trends.*
- *Produced **720 performance analytics records** with critical dropout metrics, enabling early identification of at-risk students.*
- *Achieved an impressive **~5 minute average pipeline execution time**, ensuring rapid data availability for daily insights.*
- *Maintained a **100% success rate** for all automated pipeline runs, complemented by reliable email notifications.*
- *Attained a remarkable **98.5% data quality score** post-validation, ensuring the trustworthiness of all reported data.*

Qualitative Impact:

- ***Centralized analytics platform:** Eliminated reliance on scattered Excel files, providing a single source of truth.*
- ***Automated daily reporting:** Saved an estimated **2-3 days per month** in manual data aggregation and report generation.*
- ***Early identification of at-risk schools:** Enabled proactive interventions to support schools struggling with attendance or performance.*
- ***Data-driven decision-making enabled:** Empowered educators and administrators with robust insights for strategic planning and resource allocation.*

🚀 Future Enhancements: A Phased Approach to Innovation

1: Real-Time Processing

To provide more immediate insights, the platform will evolve to support real-time data streams:

- *Implement Spark Structured Streaming for near real-time enrollment updates.*
- *Enable live dashboard refreshes, reducing latency for critical decision points.*

2: Machine Learning Integration

Leveraging advanced analytics to predict future trends and identify patterns:

- *Develop a dropout prediction model using MLflow for lifecycle management.*
- *Implement student performance forecasting to anticipate academic needs.*
- *Introduce anomaly detection for enrollment patterns to flag unusual changes.*

3: Advanced Governance

Enhancing data security, compliance, and discoverability:

- *Adopt Unity Catalog for centralized data governance and metadata management.*
- *Implement row-level security to protect sensitive student data.*
- *Establish comprehensive data lineage tracking for auditability and trust.*

4: API Layer Development

To broaden accessibility and integration capabilities:

- *Develop a REST API for programmatic access to curated data sets.*
- *Create a dedicated mobile app for key stakeholders to access insights on the go.*

Conclusion

This project successfully concludes the phase of building a foundational education analytics platform. It stands as a testament to effective data engineering practices and a commitment to data-driven insights. Looking forward, several strategic enhancements are planned to further expand its capabilities and impact.

*I have successfully built a **production-grade education analytics platform** that delivers tangible value:*

- ***Automates data processing** from raw (Bronze) to refined (Silver) and aggregated (Gold) layers.*
- ***Ensures high data quality** with a verified 98.5% completeness and accuracy score.*
- ***Delivers actionable insights** on critical metrics such as dropout risk and enrollment trends.*
- ***Demonstrates real-world data engineering practices** through robust design and implementation.*
- ***Enables data-driven education planning**, empowering stakeholders with reliable information.*