Enterprise Data Solution

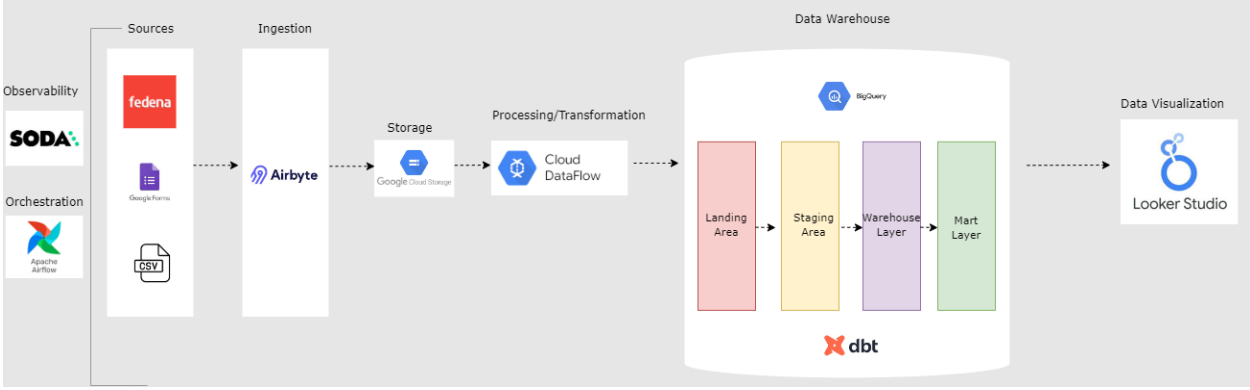
Overview

This document describes the comprehensive data infrastructure for Vine Secondary School. The goal of the infrastructure is to facilitate efficient data collection, processing, warehousing, and reporting to improve student performance, especially in exams such as JAMB and WASSCE. This solution leverages various tools to ensure data is captured from multiple sources, processed through a robust pipeline, stored securely, transformed for analytics, and visualized for informed decision-making.

The End -to- End pipeline

The solution uses the **ELT (Extract, Load, Transform)** method, where data is extracted, loaded in its raw form, and then transformed inside the data warehouse using dbt. The reason for this approach is as foolws:

1. **Flexibility in Transformation**: With raw data loaded, dbt can handle all transformations. There is flexibility in transforming and re-transforming the data as business needs evolve.
2. **Agility for Business Changes**: Since transformations happen in dbt, any changes in business logic can easily be reflected by updating the dbt models. No need to reload or restructure the raw data.
3. **Full Auditability**: The raw data is always retained in the raw data in the warehouse, which is useful for audits, debugging, or recreating historical reports with different transformations.
4. **Modularity**: dbt models are modular, which allows for the creation of reusable logic in layers (e.g., staging, transformations, analytics), making it easier to adapt to changes.



Data Collection Tools

Fedena

Fedena is a free and open-source school management software that is highly customizable and widely used in educational institutions. It offers a range of features, including attendance tracking, grade management, and timetable scheduling, all aimed at improving operational efficiency in schools.

Open Source:

Fedena is open-source, allowing institutions to adapt and customize it according to their needs. It supports both free and paid versions, but its free version is powerful enough to be used for core administrative functionalities.

Backend:

Fedena uses MySQL as its backend database, making it compatible with most data systems. Data from Fedena can be accessed directly from the MySQL database, or through various other export options like CSV and APIs.

Types of Data Collected

Student Information:

Demographic details such as name, date of birth, gender, address, parent/guardian details.

Attendance Records:

Daily student attendance, absentees, and latecomers.

Academic Data:

Grades, test results, and exam performance for each student.

Timetable Information:

Class schedules, subject allocations, and teacher assignments.

Extracurricular Activities:

Participation records for non-academic activities like sports, drama, music, etc.

Data Access Methods:

API Key: Fedena provides a secure API that can be used to pull data directly from the system. Schools can generate an API key that allows third-party tools like Airbyte (used for ingestion in this solution) to access the data programmatically.

CSV Exports: For schools that prefer manual processes, Fedena allows users to export data in CSV format. This data can then be uploaded into the storage system (GCS in this solution) for further processing.

Customization: Fedena is customizable, allowing schools to tailor their data collection forms and reporting tools to meet their specific needs. This flexibility is important for schools with unique curricula or administrative requirements.

Suitability: For Vine Secondary School, Fedena is a great fit as it allows capturing a wide range of student, teacher, and administrative data, which is crucial for building a strong foundational data infrastructure for the school's academic performance and operational insights.

Google Forms

Google Forms is a cloud-based survey and data collection tool that can be used to easily collect data from multiple users, such as teachers, students, and parents. It is part of the Google Workspace and offers easy integration with other Google tools like Google Sheets and Google Drive for storing collected data.

Types of Data Collected

Surveys: Teachers, students, and parents can fill out surveys to provide feedback on various aspects of school life, including teaching quality, exam stress levels, extracurricular activities, and more.

Forms for Feedback: Google Forms can be customized to create feedback forms for teachers to evaluate student performance or for students to give feedback on classes and subjects.

Exam Registration Forms: Google Forms can be used to collect data related to student exam registration, which can be processed and prepared for further analysis.

Data Access Methods:

Google Sheets Integration: Data collected from Google Forms is automatically stored in Google Sheets, making it easy to access, visualize, and integrate with other tools. Google Sheets can be used as a data source to feed into GCS via Airbyte.

CSV Downloads: Like Fedena, data from Google Forms can also be exported as CSV files. This allows for quick uploads into GCS for further transformation and analysis in the data pipeline.

Advantages:

Real-Time Data Collection: Responses submitted through Google Forms are updated in real-time in Google Sheets, making it a quick and efficient way to gather data from multiple users.

Ease of Use: Google Forms is very user-friendly, requiring minimal technical expertise to set up. It's a great tool for collecting survey data from students and parents.

Customization: Google Forms allows customization of forms with multiple types of questions (text, multiple-choice, drop-downs, etc.), and it is easy to set up for specific school needs.

Suitability: Google Forms is suitable for collecting feedback and survey data, making it a valuable tool for understanding issues like student stress levels, access to resources, and motivation—all of which are critical factors affecting student exam performance.

Pipelining

Ingestion

Airbyte

Airbyte is an open-source data integration tool used for connecting various data sources and ingesting data into the system. It is highly flexible and supports a wide range of data connectors.

Role in Solution:

Airbyte is responsible for ingesting data from Fedena, Google Forms, and external CSVs, and pushing it into the Google Cloud Storage (GCS) bucket.

Process:

Fedena API Integration: Using Airbyte's API connector, real-time data from Fedena can be pulled into the data pipeline.

Google Forms Connector: Airbyte can pull data from Google Sheets, which acts as the backend for Google Forms, into GCS.

CSV Uploads: For manual uploads, Airbyte will read CSV data and ingest it into the pipeline.

Why Airbyte?

Airbyte supports incremental and real-time data ingestion, making it suitable for managing the school's ongoing data needs without unnecessary data duplication.

Storage

Google Cloud Storage (GCS)

GCS is a cloud-based storage solution designed to store large volumes of data in a scalable and secure manner.

Role in Solution:

All ingested data from various sources (Fedena, Google Forms, and CSV files) is stored in GCS as the raw data landing zone.

Data is stored in its original format, ready for further transformation and processing.

Benefits:

Scalable: GCS can scale according to the needs of the school as the data volume grows.

Secure: GCS provides high levels of data encryption and security, ensuring sensitive student data is stored safely.

Processing/Transformation

Cloud Dataflow

Google Cloud Dataflow is a managed stream and batch processing tool that allows for real-time or scheduled data processing tasks.

Role in Solution:

ETL Transformations: Cloud Dataflow is used to apply any required transformations to the raw data before it is loaded into the data warehouse. This includes cleaning, formatting, and enriching the data to ensure it is ready for analytics.

Processing Large Data: It handles large volumes of data efficiently, ensuring that even as the school's data grows, the system remains performant.

Real-Time and Batch Processing: Dataflow supports both real-time stream processing and batch ETL jobs, making it a flexible tool for handling different types of data loads.

Warehousing

Google BigQuery

BigQuery is a fully-managed, serverless data warehouse designed to handle large-scale data analytics.

Role in Solution:

Data Warehousing: BigQuery serves as the central data warehouse for the school's data. Data is structured into multiple layers (Landing Area, Staging Area, Warehouse Layer, and Mart Layer) to optimize data access and query performance.

Fast Querying: BigQuery supports fast SQL queries over large datasets, allowing school administrators and stakeholders to get insights quickly.

Data Layers:

Landing Area: Raw data is ingested and stored here directly from the data sources.

Staging Area: Data is cleaned, transformed, and staged for further use.

Warehouse Layer: Data is structured and optimized for analytics.

Mart Layer: Specific, detailed data marts are created for department-specific reporting and analysis (e.g., grades, attendance, performance).

dbt (Data Build Tool)

dbt is a transformation tool that enables analysts and engineers to transform data within the warehouse using SQL.

Role in Solution:

Transformations: dbt is responsible for defining and running all transformations inside BigQuery. This includes creating the final analytics tables, data cleaning, and structuring the data into data marts for specific reporting needs.

Documentation and Testing: dbt also automatically generates documentation and tests for each transformation, ensuring data quality and transparency.

Automation

Apache Airflow

Apache Airflow is a platform to programmatically author, schedule, and monitor data pipelines.

Role in Solution:

Orchestration: Airflow manages the entire workflow of the data pipeline. It schedules and monitors each step, from data ingestion with Airbyte, to storage in GCS, processing via Dataflow, loading into BigQuery, and applying transformations using dbt.

Error Handling and Monitoring: Airflow handles retries and failures, ensuring that the pipeline is resilient and self-correcting when issues occur.

Observability

Soda

Soda is an observability tool used to monitor and improve data quality in the pipeline.

Role in Solution:

Data Quality Checks: Soda ensures that data quality is maintained at each stage of the pipeline. It monitors for data consistency, accuracy, and integrity across all sources and transformations.

Real-Time Monitoring: It provides real-time insights into data quality issues, allowing administrators to proactively address any potential issues.

Reporting

Looker Studio

Looker Studio is a free tool that transforms data into fully customizable dashboards and reports.

Role in Solution:

Data Visualization: Looker Studio is used to visualize the data stored in BigQuery. It creates interactive dashboards that allow stakeholders to view and analyze key performance metrics for the school.

Report Customization: Stakeholders can create customized reports for different use cases, such as student performance tracking, attendance analysis, and exam preparation insights.