Project Title: Identifying Data Integration Quality for Multi – Source

Analytics Team Members:

1. **Name**: GeethaPriya KA
   CAN_ID: 33738378

Institution Name: Vemana Institute of Technology

---

1. Abstract

This project investigates an AI-driven approach to identifying data integration quality in multi-source analytics. Traditional data integration methods often suffer from schema mismatches, missing values, and inconsistent formats, leading to unreliable insights. Using tools like Pandas, Scikit-learn, and visualization libraries, the system analyzes integrated datasets, detects anomalies in data consistency, and applies machine learning techniques to assess integration quality. The solution enables automated validation, scalable anomaly detection, and insightful visualizations, ensuring high-quality data for multi-source analytics and decision-making.

---

2. Problem Definition

Data integration from multiple sources is often plagued by inconsistencies, missing values, schema mismatches, and duplicate records, leading to inaccurate analytics and flawed decision-making. Traditional data validation methods are manual, time-consuming, and prone to errors, making it difficult to ensure high-quality integrated datasets for analytics.

Key Questions:

- How can AI-driven techniques improve data integration quality assessment?
- Can anomaly detection models identify inconsistencies and schema mismatches in multi-source data?
- How can automated validation be integrated into existing data pipelines?

Target Users:

- **Data Engineers:** To ensure clean and consistent data for integration.
- **Data Scientists & Analysts:** To improve the reliability of multi-source analytics.
- **Business Intelligence Teams:** To make accurate data-driven decisions based on high-quality integrated datasets.

Goal:

- Detect and flag inconsistencies in multi-source data to enhance data reliability.
- Automate integration quality checks to reduce manual validation efforts.
- Ensure high-quality data for advanced analytics and decision-making.

---

3. Requirements

Functional Requirements

- Load and process multi-source integrated data.
- Detect inconsistencies, missing values, schema mismatches, and duplicate records.
- Implement machine learning models to assess data integration quality.
- Generate reports or alerts highlighting integration anomalies.
- Provide visualizations for better understanding of data quality issues.

Non-Functional Requirements

- Ensure scalability to handle large and complex integrated datasets.
- Provide data security by maintaining confidentiality and access control.
- Ensure compatibility with existing data pipelines and analytics platforms.

---

4. Tools and Platforms

Tools:

- **Data Preprocessing:** Python, Pandas, NumPy
- **Data Quality Assessment:** Scikit-learn, PyOD (for anomaly detection)
- **Visualization:** Matplotlib, Seaborn, Plotly

Cloud Services

- **Data Storage:** AWS S3, Google Cloud Storage, or Azure Data Lake

- **Data Processing & Notebook Environment:** Jupyter Notebook, Google Colab
- **Model Training & Deployment:** IBM Watson Studio, AWS SageMaker, or Google Vertex AI.

---

5. Implementation Plan

Step 1: Data Preparation

- Upload multi-source datasets to a cloud storage platform (e.g., AWS S3, Google Cloud Storage).
- Perform data cleaning by handling missing values, schema mismatches, and duplicates.

Step 2: Model Development

- Train an Isolation Forest model using Scikit-learn to detect data integration anomalies.
- Experiment with rule-based validation and PCA-based anomaly detection.
- Evaluate and optimize the model using metrics like precision, recall, and data consistency scores.

Step 3: Deployment

- Deploy the model in IBM Watson Studio, AWS SageMaker, or Google Vertex AI as an API.
- Test the API using sample multi-source datasets.

Step 4: Reporting and Visualization

- Visualize integration inconsistencies using Matplotlib, Seaborn, or Plotly.
- Generate a data quality report highlighting integration anomalies and patterns.

---

6. Expected Outcomes

- A machine learning-based solution to assess data integration quality.
- An API for automated integration anomaly detection.
- Visual reports for data engineers and analysts to improve multi-source data reliability.