Name: Rohith kumar Senthil kumar **DS 5110**
Date: 06/08/2025 **Iteration#02 Report**

Data Fusion Pipeline

Github: github.com/Rohith-Kumar-S/datafusion-pipeline

Project Kickoff:

Goal: The goal of this project is to develop a data integration pipeline to combine data from multiple sources into a unified format.

Scope: This project should be capable of handling both structured and unstructured data in various formats, including JSON, XML, and YAML. It should also support image processing and allow transformations based on the user's requirements. Users should be able to define rules for the provided data and reuse these rules to process similar data. The system should clean and transform the data, converting it into a unified format based on the user's choice. If time permits, the project should be made scalable to handle data from big data sources, including real-time Change Data Capture (CDC) events.

Phases of deliverables & Timelines:

- Phase 1: Planning and Analysis (Weeks 1)
  Milestone: Requirements sign-off and technical architecture approval
- Phase 2: Basic data ingestion for JSON, XML, YAML formats with simple transformations (Weeks 2)
  Milestone: Development environment ready and basic data ingestion operational
- Phase 3: Add image processing capabilities starting with basic operations() (Weeks 3)
  Milestone: Core pipeline functionality complete with basic image and data processing capabilities
- Phase 4: Implement rule-based system for user-defined data processing rules (Week 4)
  Milestone: Rule-based system implemented Milestone: Data transformation capabilities ready for operations
- Phase 6: Implement unified format conversion capabilities (Week 6)
  Milestone: Core goal satisfied
- Phase 7: Add big data scalability and real-time CDC if time permits (Week 7 - 8)
  Milestone: Big data integration complete

Datasets: (To be used)
www.tablab.app/json, autonomous-driving-challenge/images

Knowledge gaps & To do's: Up skill on Big Data, reinforce current knowledge on data wrangling and image processing.

Team Discussions:

Core skills: Project Management, Data Management, Data wrangling

Roles and responsibilities:

- Research and development on designing a pipeline architecture for data ingestion, processing, storage and accessing.
- Building the whole application end-to-end
- Quality Assurance, testing and validation
- Documentation of methodologies and results

Skill gaps: Big data

Programming language: Python

Skills and Tools Assessment:

Tools to be used:
Data Science Tools: Pandas, numpy, matplotlib, seaborn, streamlit
Computer vision tools: OpenCV
DataBase: MySQL

Submission for This Iteration:

Excel tracker is uploaded to the github repository.