ETL Pipeline - Technical Documentation

1. Overview

This document outlines the design, implementation, and CI/CD integration of an ETL Pipeline project for stock market data enrichment and storage using MongoDB.

2. Pipeline Design

The pipeline performs the following steps:

- Extracts stock data from a CSV file.
- Enriches it using real-time APIs (Yahoo Finance, NewsAPI, Finnhub, and ExchangeRate API).
- Transforms data with unit conversions, feature engineering, and timestamp formatting.
- Loads the cleaned data into MongoDB.

Design ensures modularity and reusability through well-defined enrichment functions.

3. Technology Choices

- Language: Python (easy syntax, rich data libraries like pandas, requests).
- Scheduler: schedule module (lightweight, readable for simple job scheduling).
- Database: MongoDB (flexible schema, suitable for JSON-like enriched data).
- Deployment: GitHub Actions (for CI/CD automation).

4. CI/CD Integration

GitHub Actions automates the testing and deployment process.

Steps included:

- Run automated unit tests on each push/merge.
- Validate data schema and pipeline functionality.
- Deploy scripts to the production environment.

Benefits of CI/CD:

- Reduces manual errors by automating builds and tests.
- Facilitates rapid feedback during development.
- Ensures data integrity through continuous validation.
- Accelerates development cycles.

5. Future Enhancements

- Add retry logic for API failures.
- Store historical API responses for traceability.
- Monitor pipeline health and errors using alerting tools.