Technical Documentation

CSV Parser Module

---

1. Introduction

The CSV Parser Module is a core component of the Data Ingestion Pipeline for Project Atlas. Its primary role is to process client-provided CSV files containing sensitive financial data and prepare them for downstream validation and analytics. The module emphasizes correctness, compliance, and robustness while integrating seamlessly with the existing legacy validation framework.

---

2. Objectives

- Parse client CSV files with varying schemas and formats
- Ensure data integrity and regulatory compliance
- Integrate with the legacy validation framework without modifying it
- Handle malformed or invalid data gracefully
- Log errors and anomalies using predefined logging standards

---

3. System Context

The CSV Parser Module sits between the file ingestion layer (AWS S3) and the legacy validation framework. Files are retrieved from S3, parsed by this module, and then passed to the validation layer for compliance checks before entering downstream analytics systems.

---

4. Architecture Overview

Input: Raw CSV files from AWS S3
Processing:

- Schema detection and flexible column handling
- Row-by-row parsing
- Type conversion and basic sanity checks
- Error detection and logging

Output: Structured data objects compatible with the legacy validation framework

---

5. Functional Design

5.1 CSV Parsing Logic

- Uses Python-based CSV parsing utilities
- Supports optional and additional columns beyond the core schema
- Handles delimiter inconsistencies and missing headers

5.2 Error Handling

- Malformed rows are skipped by default
- Errors are logged with context (file name, row number, error type)
- Processing continues without terminating the entire job

5.3 Validation Integration

- Parsed records are passed to the legacy validation adapter
- Validation results determine whether records proceed or are rejected
- No changes are made to existing validation logic

---

6. Compliance Considerations

- No sensitive data is written to logs
- Processing occurs only within approved infrastructure
- Audit-friendly logging is enabled for traceability
- Non-compliant records are flagged and isolated

---

7. Assumptions

- Some CSV schemas may evolve over time
- Large file performance optimization is deferred
- Invalid rows are non-critical and can be skipped

---

8. Limitations

- Not optimized for very large files (>10GB)
- Depends on finalized schema definitions for full validation
- Relies on staging/local datasets for testing

9. Testing Strategy

- Unit tests validate parsing logic and edge cases
- Test coverage target: minimum 80%
- Includes tests for malformed rows and missing fields

---

10. Conclusion

The CSV Parser Module delivers a reliable and compliant solution for ingesting financial CSV data into Project Atlas. Its modular design, strong error handling, and seamless legacy integration ensure maintainability and audit readiness while allowing future performance enhancements.