Project Atlas – Data Validation Module

Technical Specification (Week 1 Analysis)

1. Overview

Technical specifications for the **Data Validation Module** within Project Atlas, a **Customer Data Integration Platform** for a financial services client. The purpose of this module is to ensure data quality, regulatory compliance (GDPR & CCPA), and schema integrity before data enters downstream ETL and analytics workflows.

The validation module will act as a **pre-transformation gate** in the existing AWS Glue / Apache Spark ingestion pipeline. Records failing validation will be rejected or flagged according to defined severity rules, preventing downstream processing errors and compliance risks.

2. Scope and Objectives

In Scope:

- Schema validation against predefined data structures
- Data type enforcement
- Basic business rule validation (required fields, ranges, nullability)
- Validation logging for compliance and monitoring

Out of Scope:

- Full implementation of validation logic
- Performance benchmarking and tuning
- UI or dashboard changes
- Advanced error remediation workflows

3. Data Source Overview

Dataset

- **Name:** user_behavior_events.json
- **Purpose:** Captures user interaction events from the client's financial services platform
- **Volume:** 10,000 records

- **Noise Level:** 5%

Core Fields

- user_id (UUID)
- action (string)
- page (string)
- duration_ms (integer, nullable)
- device_type (string)
- timestamp (datetime, ISO format)
- referrer (string, nullable)
- session_id (string)
- ip_address (string, nullable)

4. Validation Rule Categories

Schema Validation

Key checks include:

- Presence of all mandatory fields
- No unexpected or extra fields
- Correct nesting and field naming

Mandatory fields:

- user_id
- action
- page
- device_type
- timestamp
- session_id

Data Type Validation

Each field must conform to its defined data type.

Examples:

- user_id must be a valid UUID
- duration_ms must be an integer when present
- timestamp must be parseable as a valid datetime
- Nullable fields (referrer, ip_address) must allow null values

Invalid type records will be rejected.

5. Compliance Considerations (GDPR & CCPA)

Key considerations:

- **Data minimization:** Only required fields are processed
- **PII handling:** ip_address treated as sensitive data
- **Logging controls:** Validation logs must avoid storing raw IP addresses
- **Auditability:** All validation failures must be traceable

No external APIs or third-party validation services will be used due to compliance constraints.

6. Error Handling Strategy

- Invalid records are removed from the main data flow
- Validation errors are logged with:
    - Rule violated
    - Field name
    - Severity level
- Downstream ETL processes receive only validated records

Exact remediation or reprocessing workflows are pending client clarification.

7. Integration with Existing Pipeline

The Data Validation Module will:

- Run within the existing AWS Glue / Spark environment
- Execute **before transformation stages**
- Output:
    - Valid dataset → downstream ETL
    - Validation logs → compliance monitoring tools

8. Assumptions

- Full schema documentation for legacy sources will be provided
- Validation rules documented in knowledge base are mostly accurate
- Infrastructure stability during development and testing
- Basic logging-based error handling is acceptable for this phase

9. Risks and Open Questions

Identified Risks

- Incomplete or undocumented legacy schemas
- Ambiguity in handling partially valid records
- Timestamp format inconsistencies across sources
- Potential data loss due to strict validation rules

Open Questions

- Should warnings be allowed into downstream systems?
- How long should validation logs be retained?
- Are schema versions expected to evolve?

10. Conclusion

This document provides a structure of the Data Validation Module for Project Atlas. It provides clear validation categories, compliance considerations, and integration points, forming the foundation for the implementation phase.