Integration Specification

Feature Extraction Module – Project Atlas (Week 2)

1. Document Overview

This document defines how the Feature Extraction Module integrates with the existing Analytics Engine data processing pipeline, including upstream inputs, downstream outputs, and validation touchpoints.

2. Integration Objective

The objective of this integration is to:

- Seamlessly introduce unstructured text processing into the existing pipeline
- Preserve all upstream ETL and downstream analysis components unchanged
- Ensure schema compatibility and governance compliance
- Minimize performance impact on the shared Spark cluster

3. Integration Scope

In Scope

- Integration of the Feature Extraction Module into the existing feature extraction stage
- Consumption of Spark DataFrames produced by the ETL layer
- Emission of structured features for downstream analytics

Out of Scope

- Changes to upstream ETL logic
- Modifications to downstream analytics components
- Database schema changes
- UI or reporting layer changes

4. Pipeline Integration Point

Current Pipeline Flow

Legacy S3 Buckets

ETL & Validation Layer

Feature Extraction Stage

Downstream Analytics

Updated Flow (With This Module)

Legacy S3 Buckets

ETL & Validation Layer

Feature Extraction Stage

Structured Feature Extractor (existing)

Unstructured Feature Extractor (this module)

Downstream Analytics

The new module operates **in parallel** with existing structured feature extraction logic.

5. Upstream Integration

Input Source

- Apache Spark DataFrame produced by the ETL and validation layer
- Data originates from legacy S3 buckets

Required Input Contract

- Spark DataFrame must contain the following column:
  - text (string): unstructured text data (logs, notes)

Input Assumptions

- Text may be empty or malformed
- Additional columns may exist but are ignored
- Schema documentation may be partial; core fields are assumed stable

Validation Hook

- Module performs input validation:
  - Confirms presence of text column
  - Raises controlled exception if validation fails

6. Internal Integration Logic

Execution Model

- Module is invoked as part of the feature extraction stage
- Operates on Spark DataFrames using columnar transformations
- No Spark UDFs or external dependencies

Processing Steps

1. Validate input schema
2. Sanitize text to remove PII indicators
3. Extract structured features
4. Emit schema-compatible DataFrame

7. Downstream Integration

Output Contract

- Apache Spark DataFrame
- Structured schema aligned with downstream analytics expectations

Output Schema

Column Name	Type	Description
text	string	Original text
text_length	int	Length of sanitized text
word_count	int	Number of tokens
keyword_density	float	Word count / text length
empty_text_flag	int	Empty text indicator

Downstream Compatibility

- No schema migration required
- No changes required in downstream analysis components
- Output conforms to existing structured feature ingestion logic

8. Validation and Governance Integration

Governance Hooks

- Sanitization logic removes common PII patterns
- No raw PII features are generated
- Processing occurs entirely within internal Spark cluster

Validation Layer Interaction

- Existing pipeline validation checks remain unchanged
- Module outputs pass through standard validation harness

9. Error Handling & Failure Modes

Error Scenarios

- Missing required text column
- Invalid DataFrame schema

Handling Strategy

- Raise controlled, descriptive exceptions
- Fail fast at feature extraction stage
- Prevent propagation of malformed data downstream

10. Performance Considerations

- Designed for lightweight execution
- Columnar Spark operations only
- Minimal memory footprint
- Suitable for shared cluster execution
- No impact on upstream or downstream SLAs

11. Deployment & Enablement

Deployment Model

- Deployed alongside existing feature extraction components
- Enabled via pipeline configuration (no code changes upstream/downstream)

Rollback Strategy

- Module can be disabled without affecting existing pipeline
- Structured feature extraction remains unaffected

12. Assumptions & Constraints

Assumptions

- Legacy schema remains stable for core fields
- Unstructured data volume is moderate for prototype phase
- Baseline performance metrics are estimated

Constraints

- Shared cluster resources
- Python + Apache Spark only
- No external APIs or services
- Hard Week 2 deadline

13. Conclusion

This integration approach ensures that the Feature Extraction Module can be introduced into the Analytics Engine with minimal risk, no breaking changes, and full compliance with governance and performance constraints. The design prioritizes isolation, compatibility, and ease of review by senior engineers.