Compliance & Data Governance Checklist

Feature Extraction Module – Project Atlas (Week 2)

1. Document Overview

This document outlines the data **governance and compliance** measures implemented in the **Feature Extraction Module** to ensure adherence to policies, privacy requirements, and shared infrastructure constraints.

2. Governance Objectives

The primary objectives of this checklist are to:

- Prevent exposure of Personally Identifiable Information (PII)
- Ensure secure and compliant data processing
- Restrict usage to approved internal infrastructure
- Maintain auditability and review readiness

3. Data Handling & Privacy Controls

3.1 PII Identification

PII Category	Potential Presence	Mitigation Strategy
Email addresses	Possible	Removed via regex sanitization
Numeric identifiers (IDs)	Possible	Removed during text sanitization
Phone numbers	Possible	Excluded via numeric pattern filtering
Names	Possible	Not extracted as features
IP addresses	Possible	Not extracted or stored

3.2 Text Sanitization Measures

- All text converted to lowercase
- Email patterns removed
- Numeric identifiers removed
- Sanitized text used only for feature computation
- Raw text not transformed into derived PII signals

4. Data Minimization

- Only essential features are extracted
- No raw text fields are duplicated or expanded
- No unnecessary intermediate data persisted
- Output limited to analytical signals only

5. Infrastructure & Processing Constraints

5.1 Processing Environment

Requirement	Status
Internal Spark cluster only	Compliant
Shared cluster usage	Compliant
No external SaaS tools	Compliant
No cloud-based NLP services	Compliant

5.2 Dependency Management

- Only standard Python libraries used
- Apache Spark APIs only
- No third-party NLP or ML frameworks
- No internet or external API access

6. Access Control & Data Exposure

- Module does not persist data externally
- No data written to external storage
- Access governed by existing pipeline permissions
- No elevation of privileges required

7. Logging & Observability

- No logging of raw text content
- Error messages do not expose input data

8. Validation & Audit Readiness

- Clear documentation of sanitization logic
- Deterministic feature extraction logic
- Unit tests validate PII removal
- Design decisions documented for review

9. Regulatory Alignment

Applicable Regulations

Regulation	Applicability	Compliance
GDPR	Potential	Compliant
CCPA	Potential	Compliant
Internal Policies	Applicable	Compliant

10. Risk Assessment

Risk	Impact	Mitigation
Accidental PII leakage	High	Regex-based sanitization
Schema changes	Medium	Input validation
Performance overuse	Medium	Lightweight Spark ops

11. Compliance Sign-Off

- PII exposure risk evaluated and mitigated
- Infrastructure and dependency constraints met
- Documentation and testing complete

12. Conclusion

The Feature Extraction Module complies with data governance policies and relevant privacy regulations. It is safe for deployment within the shared Spark cluster and suitable for integration into the existing analytics pipeline.