Data Governance Policy – Spotify

1. Organizational Governance

1.1. Governance Structure

Spotify organizes its data governance across three complementary levels, ensuring strategic steering, tactical management, and documented operational execution as outlined in the organizational charter (see annex).

1.2. Legal Collaboration

The legal team and Data Protection Officer (DPO) collaborate to ensure continuous compliance with legal obligations, manage regulatory risks, and engage with regulators.

2. Data Classification, Protection, and Management

2.1. Classification by Sensitivity and Handling

- **Red Level (Sensitive Data)**: User identification data (name, email, phone number, stored credit cards, payment history, precise location).
  Measures: AES-256 encryption (at rest and in transit), pseudonymization, strict least-privilege access, 12-month access logging.
- **Orange Level (Operational Data)**: Music preferences, interaction logs (playtime, playlists, user activities, subscription types).
  Measures: Anonymization after 6 months, role-based access control (RBAC), tagging via Collibra central catalog.
- **Green Level (Public Data)**: Artist metadata, reference data, catalogs.
  Measures: Rigorous quality checks, open internal access.

2.2. Data Lifecycle and Accessibility

- Databricks to unify data operations across all the organization
- Collibra implementation to centralize data documentation and ensure secure, role-based access.
- Defined lifecycle stages (creation, archiving, deletion) to prevent redundancies and ensure compliance.

3. Data Infrastructure and Processing Workflows

3.1. Data Ingestion & Storage

- Raw data collection via Spotify APIs or log files into AWS S3 (Data Lake).
- Structured storage (MDM/RDM) in AWS Redshift (Data Warehouse) after transformation.
- ETL/orchestration via Apache Airflow and Python for extraction, transformation, and loading.

3.2. Governance, Quality & Security

- **Governance**: Collibra for cataloging, ownership, and metadata management.
- **Quality**: Informatica Data Quality for validation, deduplication, and business rules.
- **Security**: Splunk for monitoring, alerts, and access logging.
- **Documentation**: GitHub for version control.
- **Compliance**: VeraSafe/TrustArc for GDPR audits and user consent management.
- **APIs**: Spotify Web API (via AWS Gateway) for metadata and audio feature retrieval.

3.3. Processing & Manipulation

- Distributed processing via Apache Spark (PySpark, Spark SQL) for large-scale transformations.
- Local feature engineering using Python (pandas, numpy).

3.4. Modeling & AI

- Model training with Python, TensorFlow, PyTorch, and Scikit-learn for recommendations, classification, and clustering.
- External web scraping (if required) via Python (BeautifulSoup).
- AI bias management using TensorFlow Fairness.

3.5. Exploration & Analysis

- **BI**: Tableau for dashboards and data visualization
- **CRM**: Salesforce Media Cloud for customer segmentation and interactions.

4. Security and Compliance

4.1. Technical Measures

- Zero Trust Architecture with Okta.
- RBAC via Splunk (roles: business reps → processed data; analysts → aggregated data; engineers → pseudonymized raw data; DPO/legal → full audit access).
- Mandatory 2FA across all systems (Spotify email + Splunk authentication).
- CloudFlare Firewall and VirusTotal antivirus with weekly vulnerability scans, quarterly pentests, and annual third-party audits.
- ISO 27001 and PCI-DSS compliance (via Stripe for payments).
- SIEM (Splunk) with exhaustive activity logging.
- **Disaster Recovery**: Daily AWS S3 backups.
- **API Security**: AWS API Gateway.
- **Encryption**: AES-256 for Red/Orange level data.
- **Training**: Clear security policies and phishing simulations.

4.2. Regulatory Compliance & GDPR

- VeraSafe (GDPR/CCPA audits) and TrustArc (privacy management).
- Strict adherence to user rights (access, rectification, erasure, portability).
- Granular consent management tools.
- 72-hour breach notification.

5. Data Quality and Management

5.1. Quality Assurance

- KPIs via Informatica dashboards to measure completeness, accuracy, consistency, and integrity.
- Data Stewards/Owners responsible for anomaly resolution.
- Metadata cataloging in Collibra for traceability.

5.2. Architecture & Integration

- **MDM**: Unified master data (users, content, transactions) stored in AWS Redshift.
- **Relational Databases**: Microsoft SQL Server.
- **APIs**: Cross-tool integration to eliminate silos.
- **Data Lakes**: AWS S3 for unstructured data (IoT, social media, public catalogs).

6. Transparency, User Rights, and Ethics

6.1. Transparency

- Dedicated user portal for data access, consent management, and algorithm explanations.
- Internal dashboards for data quality and bias monitoring.

6.2. User Rights

- Explicit opt-in for new features.
- GDPR/CCPA rights mechanisms embedded in systems.
- User control over data collection and deletion (opt-out).

6.3. Ethical Framework

- Independent ethics committee overseeing algorithm development.
- Pre-deployment fairness testing.
- Data minimization (retention limited to 2 years).

7. Data Culture and Continuous Improvement

7.1. Culture & Training

- Modular data literacy programs (technical, product, marketing roles).
- Workshops, assessments, and internal ambassadors.
- KPIs for practice adoption.

7.2. Monitoring & Evolution

- Annual data maturity review (benchmarked against tech leaders).
- Bug Bounty programs and user feedback surveys.

Key Objectives:

Domain	KPI	Target
Security	Major incidents	0/year
Quality	Data errors	<0.5%
Compliance	GDPR response time	≤10 days

Conclusion

This Data Governance Policy ensures robust, secure, and compliant data management at Spotify while fostering innovation and user trust through transparency, ethical responsibility, and performance. Implementation will be overseen by a dedicated governance committee to adapt to evolving technological and regulatory landscapes.

Approved by the Governance Committee on April 16, 2025