☑ 15. Metadata Management & Data Security — 50 respuestas técnicas

---

1. ¿Qué es metadata y qué tipos existen?

Metadata = *datos sobre los datos*.
Tipos:

- **Técnica:** esquemas, tipos, columnas, particiones, formato.
- **De negocio:** definiciones, KPIs, reglas.
- **Operacional:** logs, frecuencia de carga, estados, tiempo de procesamiento.

---

2. ¿Qué es metadata management y por qué es importante?

Proceso de **capturar, organizar, documentar y gobernar metadatos** para mejorar:

- Calidad
- Trazabilidad
- Descubrimiento
- Cumplimiento normativo
- Reutilización de datos

---

3. Metadatos activos vs pasivos

- **Activos:** se actualizan automáticamente (lineage automático, monitoreo).
- **Pasivos:** documentación estática manual.

---

4. Data catalog

Herramienta que centraliza metadatos, linaje, definiciones y accesos.
Facilita **data discovery**, gobernanza y calidad.

---

5. Data lineage

Trazabilidad del recorrido del dato: origen → transformaciones → destino.
Es un **subtipo de metadata técnica y operacional**.

---

6. Glossary de negocios

Catálogo de definiciones de negocio (KPI, métricas).
Deben mantenerse con:

- Owners
- Workflow de aprobación
- Versionado
- Políticas de stewardship

---

7. Diccionario de datos vs catálogo de datos

- **Data dictionary:** estructura técnica (columnas, tipos, constraints).
- **Data catalog:** incluye diccionario + linaje + calidad + gobierno.

---

8. Herramientas para metadata management

- **Cloud:** Google Data Catalog, Glue, Purview, Unity Catalog.
- **Enterprise:** Collibra, Alation, Informatica EDC.
- **Open source:** Amundsen, DataHub, Marquez.

---

9. Automatizar captura de metadatos

- Connectores nativos
- Hooks en pipelines
- APIs del catalog
- Lineage automático con Spark/dbt
- ETL instrumentado

---

10. Data profiling

Análisis de patrones, distribución, nulls, cardinalidad.
Aporta metadata de calidad.

---

11. Data dictionary y gobernanza

Proporciona claridad técnica para stewards, auditores y desarrolladores.

---

12. Business rules

Reglas que definen validez del dato (ej: “edad ≥ 18”).

---

13. Technical metadata

Schemas, modelos, formatos, tipos de datos, particiones, rutas de archivos.

---

14. Operational metadata

Logs, runtimes, fallos, timestamp, volumen procesado.

---

15. Trazabilidad de cambios

Versioning + audit logs + time travel + control de esquema.

---

16. Metadata repository

Base central donde se almacenan todos los metadatos.

---

17. Metadata federation vs central repository

- **Federation:** catálogo virtual sin mover metadatos.
- **Central:** todo se almacena en un repositorio único.

---

18. Data discovery

Proceso para explorar y encontrar datasets relevantes basado en metadatos.

---

19. Versionar metadatos

- Git (glossary, reglas)
- Versiones de schema
- Data catalog con historial
- Control de cambios automatizado

---

20. Auditar cambios en metadata

Audit logs + workflow + versionado + accesos registrados.

---

21. Data security

Conjunto de políticas y controles para proteger datos contra accesos indebidos.

---

22. Seguridad a nivel acceso vs columna

- **Acceso:** permisos por tabla.
- **Columna:** masking, redacción o filtrado por atributo.

---

23. Data masking

Ocultar datos sensibles (regex replacement, hashing, nulling).

---

24. Tokenization

Reemplazar valores reales por tokens irreversibles que preservan formato.

---

25. Encryption at rest / in transit

- **At rest:** cifrado en disco (AES-256).
- **In transit:** TLS/SSL.

---

26. Claves gestionadas por proveedor vs cliente

- **Proveedor (KMS-managed):** automático, fácil.
- **Cliente (CMEK/CSK):** control total, auditoría estricta.

---

27. RBAC

Roles → permisos.
Usuarios asignados a roles.

---

28. ABAC

Acceso según atributos (usuario, dataset, contexto).

---

29. Auditar accesos a datos sensibles

- Logs de lectura
- IAM audit
- Alertas por anomalías
- SIEM

---

30. Data anonymization

Elimina posibilidad de reversión (k-anonymity, differential privacy).

---

31. Anonimización vs pseudonimización

- **Anonimización:** irreversible.
- **Pseudonimización:** reversible con clave.

---

32. Data governance aplicado a seguridad

Políticas + roles + certificación + cumplimiento.

---

33. Manejo de credenciales en pipelines

- Secret managers
- Rotate keys
- No hardcode
- IAM roles temporales

---

34. Monitoreo actividad sospechosa

Anomalías de lectura, volumen, horarios inusuales, acceso repetido.

---

35. Data classification

Categorizar datos: público, interno, confidencial, sensible.

---

36. Regulaciones relevantes

GDPR, HIPAA, SOC2, ISO 27001, PCI DSS, CCPA.

---

37. GDPR vs HIPAA vs CCPA

- **GDPR:** privacidad global, consentimiento estricto.
- **HIPAA:** datos de salud en EE. UU.
- **CCPA:** privacidad de consumidores en California.

---

38. Least privilege principle

Asignar el permiso mínimo necesario.
Aplicación: RBAC/ABAC, IAM, políticas de expiración.

---

39. Auditorías externas

- Evidencias
- Lineage
- Logs
- Controles de acceso
- Políticas de retención
- Cumplimiento de estándares

---

40. Data retention policy

Define cuánto tiempo mantener datos según regulaciones y negocio.

---

41. Data lifecycle management

Seguridad en cada etapa: creación → uso → archivo → eliminación.

---

42. Proteger metadatos en cloud

- Encryption
- IAM
- Seguridad en catálogos
- Auditoría
- Control de acceso a schemas

---

43. Sensitive data discovery

Identificación automática de PII/PHI con clasificadores (regex, ML).

---

44. Data access monitoring

Seguimiento continuo de quién accede, con qué frecuencia y desde dónde.

---

45. Data breach

Exposición de datos sensibles.
Mitigación:

- DLP
- Encryption
- Tokenization
- Alertas
- Zero Trust

---

46. Seguridad en pipelines

- Secret manager
- IAM estrictos
- Least privilege
- Escaneo de código
- Validaciones de datos

---

47. Security policies as code

Seguridad declarativa en repositorios (OPA, Sentinel, IAM templates).

---

48. Balance seguridad vs accesibilidad

- RBAC + ABAC
- Data sandboxing
- Versionado
- Dashboards certificados
- Data products

---

49. Buenas prácticas de logging

- Centralizar logs
- Retención segura
- Correlación con SIEM
- Enmascarar datos sensibles

---

50. Programa integral de metadata management + data security

1. Definir owners
2. Crear catálogo central
3. Lineage automático
4. Calidad y reglas
5. Security by design
6. Clasificación de datos
7. Access monitoring
8. Auditoría
9. Mejora continua
10. Integración con CI/CD