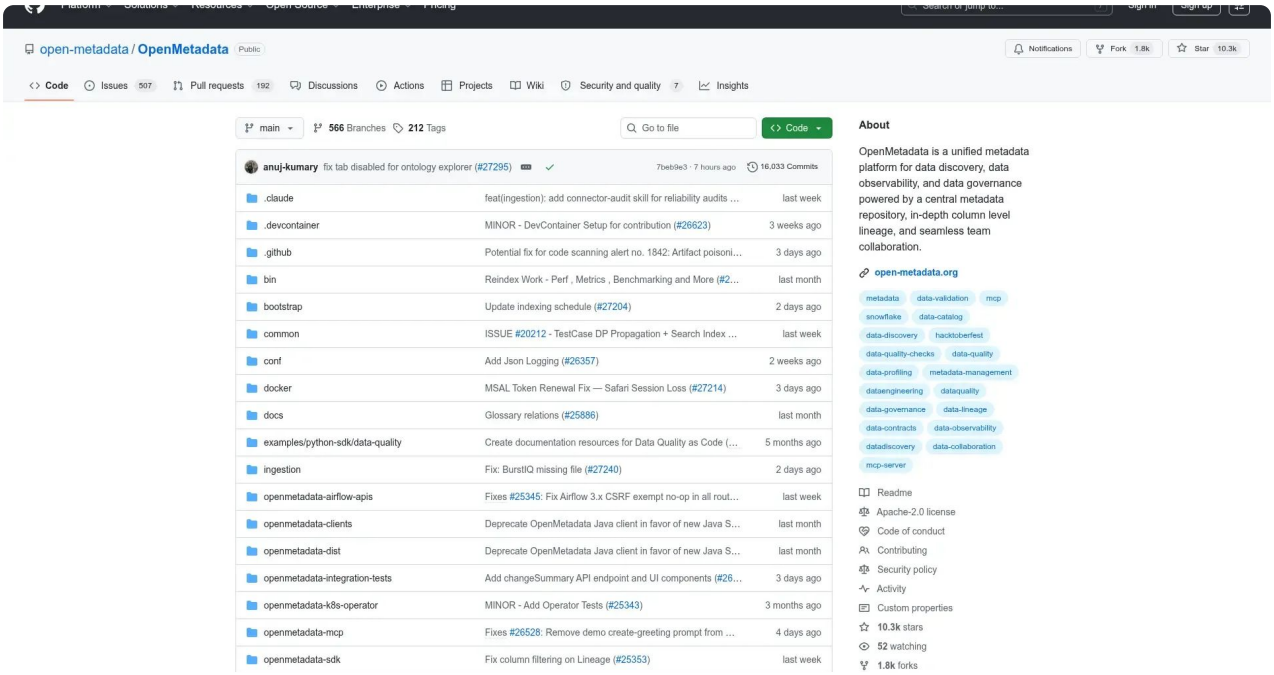

OpenMetadata

Plataforma unificada de código abierto diseñada para la gestión integral de metadatos, abarcando el descubrimiento de datos, la observabilidad y el gobierno bajo un estándar común. Está dirigida a organizaciones que buscan centralizar el conocimiento sobre su ecosistema de datos, facilitando la colaboración entre equipos técnicos y de negocio. Es ideal para perfiles como Data Engineers, Data Stewards y Analistas de Datos que operan en entornos complejos con múltiples fuentes de información.

Visitar Sitio Oficial Preguntar a ChatGPT Preguntar a Claude Preguntar a Grok

Contenido del Dossier

INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es
OpenMetadata es una plataforma unificada de código abierto diseñada para la gestión integral de metadatos, abarcando el descubrimiento de datos, la observabilidad y el gobierno bajo un estándar común. Está dirigida a organizaciones que buscan centralizar el conocimiento sobre su ecosistema de datos, facilitando la colaboración entre equipos técnicos y de negocio. Es ideal para perfiles como Data Engineers, Data Stewards y Analistas de Datos que operan en entornos complejos con múltiples fuentes de información (bases de datos, dashboards, pipelines).

Principal ventaja profesional
Permite establecer una "fuente única de verdad" totalmente automatizada gracias a su motor de ingesta y linaje a nivel de columna, eliminando los silos de información y garantizando que cualquier profesional de la empresa pueda entender el origen, la calidad y el propósito de los datos sin depender de consultas manuales constantes.

Para quién no es
No es una solución adecuada para pequeñas empresas con una infraestructura de datos mínima (una única base de datos y un reporte sencillo) o para equipos que no tienen capacidad técnica para el despliegue y mantenimiento de infraestructuras basadas en Docker o Kubernetes. Tampoco es para organizaciones que no priorizan el gobierno de datos o que buscan una herramienta puramente de visualización de negocio.

funcionalidades clave
- **Descubrimiento de Datos**: Buscador avanzado para localizar tablas, tópicos de mensajería, dashboards y pipelines mediante etiquetas, términos de glosario o lenguaje natural.
- **Linaje de Datos de Extremo a Extremo**: Visualización automática y editable del flujo de datos, permitiendo rastrear transformaciones incluso a nivel de columna individual.
- **Observabilidad y Calidad**: Implementación de pruebas de datos sin código (no-code) y perfilado automático para monitorizar la salud y fiabilidad de la información.
- **Gobierno y Glosario de Negocio**: Definición de vocabularios comunes, políticas de acceso basadas en roles (RBAC) y flujos de aprobación para cambios en los metadatos.
- **Colaboración Nativa**: Sistema de hilos de conversación, anuncios, alertas y tareas integradas directamente sobre los activos de datos.
- **Versionado de Metadatos**: Registro histórico de todos los cambios realizados en las estructuras de datos, permitiendo auditoría y seguimiento de evoluciones.

Precios
- **Versión Gratuita**: Open Source (Licencia Apache 2.0). Es la versión completa y funcional disponible en su repositorio, sin costes de licencia pero requiere gestión de infraestructura propia.
- **Versión de Pago (Collate)**: Es la oferta SaaS/Managed de los creadores de la herramienta.
- **Cloud / Single Tenant**: Precios bajo presupuesto (estimados habitualmente en escala empresarial). Incluye soporte garantizado, AI Studio, integraciones avanzadas como GitHub Metadata Sink y hosting gestionado.

Perfil del usuario
- **Empresas**: Organizaciones medianas y grandes con arquitecturas de datos distribuidas, entornos multi-cloud o necesidades estrictas de cumplimiento normativo (GDPR).
- **Perfiles Profesionales**:
- Data Engineers (mantenimiento y automatización).
- Data Stewards (gobierno y calidad).
- Data Analysts y Scientists (descubrimiento y validación).
- Chief Data Officers (CDO) (visibilidad estratégica y cumplimiento).

Nivel técnico requerido
- **Para uso**: Bajo-Medio. La interfaz de usuario es intuitiva y permite realizar la mayoría de las tareas de gobierno y calidad sin escribir código.
- **Para instalación/configuración**: Alto. Requiere conocimientos sólidos en Docker, Kubernetes (Helm Charts), gestión de servicios como Elasticsearch/OpenSearch, MySQL/PostgreSQL y configuración de red/seguridad (SSO, OAuth).

- **Necesidades de soporte**: El departamento de infraestructura o DevOps será necesario para el despliegue inicial y el mantenimiento de los servicios subyacentes.

Ejemplos de uso profesional
- **Impact Analysis**: Evaluar instantáneamente qué informes de PowerBI o Tableau se verán afectados si se modifica el nombre de una columna en el Data Warehouse.
- **Certificación de Datos**: Marcar activos de datos como "Tier 1" o "Verificados" para que los analistas de negocio sepan qué fuentes son oficiales y fiables.
- **Automatización de PII**: Detectar automáticamente columnas con información sensible (DNI, tarjetas, emails) y aplicar etiquetas de privacidad para cumplir con regulaciones.

Uso y distribución
- **Versión web**: Interfaz principal accesible mediante navegador tras el despliegue.
- **Versión escritorio**: No dispone de aplicación nativa (orientado a entorno servidor).
- **Versión móvil**: Interfaz web responsiva.
- **CLI**: Herramienta de línea de comandos en Python para automatizar ingestas y auditorías.
- **SDK**: Librerías oficiales para Python, Java y TypeScript para desarrollo de integraciones personalizadas.

Open source
Proyecto bajo Licencia Apache 2.0, con más de 10.000 estrellas en GitHub y una comunidad muy activa que publica actualizaciones mensuales de forma cadencial.

Integraciones
- **Facilidad de integración**: De No-code (conectores UI) a Full-code (APIs y SDKs).
- **API propia**: API REST extensiva basada en estándares OpenMetadata para gestionar cualquier entidad mediante código.
- **Servidor MCP**: Integración reciente que permite a agentes de IA interactuar con el catálogo de datos mediante el protocolo Model Context Protocol.
- **Integraciones Nativas**: Más de 84 conectores preconstruidos que incluyen Snowflake, BigQuery, Redshift, Databricks, dbt, Airflow, Glue, Kafka, Tableau, PowerBI y Looker.
- **Alertas**: Integración con Slack, Microsoft Teams y Google Chat a través de Webhooks para notificaciones en tiempo real sobre cambios o fallos de calidad.

Notas finales
información legal, licencias, contratos
El software se distribuye "tal cual" bajo la licencia Apache 2.0. El usuario mantiene la propiedad total de sus metadatos. En el caso de optar por Collate (versión gestionada), se aplican contratos de servicio (SLA) comerciales y condiciones de seguridad específicas del proveedor.

Otros
OpenMetadata destaca por su enfoque en estándares (JSON Schemas) para definir la semántica de los metadatos, lo que evita el "vendor lock-in" o dependencia exclusiva de la herramienta para acceder a la información gestionada.

Para más información:
- Sitio web oficial: https://open-metadata.org
- Precios (Collate): https://www.getcollate.io/pricing
- Documentación técnica: https://docs.open-metadata.org
- Github: https://github.com/open-metadata/OpenMetadata
- Slack de la comunidad: https://slack.open-metadata.org_

CONSEJOS DE IMPLANTACIÓN

Aplicación profesional
Empresas de mediano y gran tamaño con arquitecturas de datos complejas que requieren centralizar el gobierno, la observabilidad y el linaje. El presupuesto para la versión Open Source es de cero euros en licencias, pero requiere inversión en infraestructura (Cloud/On-premise) y horas de ingeniería. Para la versión gestionada (Collate), el presupuesto se sitúa en rangos empresariales (SaaS). Los puntos clave incluyen la eliminación de silos de información, el cumplimiento normativo (GDPR/PII) y la reducción del tiempo de descubrimiento de activos de datos.

Madurez digital requerida
- **Usuarios y equipo**: Los usuarios finales del negocio requieren una madurez básica en el consumo de datos, mientras que los equipos técnicos (Data Engineers/Analysts) deben estar familiarizados con conceptos de metadatos, calidad de datos y SQL.
- **Empresa y departamentos**: La organización debe contar con una estructura de datos ya establecida (Data Warehouse, Data Lake o múltiples bases de datos) y una cultura que valore la gobernanza. Es indispensable contar con un departamento de ingeniería o DevOps con capacidad para gestionar contenedores y orquestación.

Plan orientativo de implantación
Pasos necesarios y estimaciones
- **Tiempos estimados de despliegue**: De 4 a 12 semanas para una fase inicial funcional, dependiendo de la cantidad de fuentes de datos.
- **Evaluación inicial (1-2 semanas)**: Inventario de fuentes de datos (bases de datos, BI, Pipelines), definición de roles de acceso y selección de casos de uso críticos (ej. linaje de reportes financieros).
- **Implantación inicial y PoC (2-3 semanas)**: Despliegue de la infraestructura OpenMetadata en Kubernetes o Docker. Configuración de la base de datos de metadatos (MySQL/PostgreSQL) y el motor de búsqueda (OpenSearch/Elasticsearch).
- **Configuración e ingesta (2-4 semanas)**: Conexión de las primeras fuentes (Snowflake, BigQuery, etc.), configuración de perfiles de ingesta y mapeo inicial del Glosario de Negocio.
- **Formación y capacitación (1-2 semanas)**: Talleres prácticos para Data Stewards y Analistas sobre el uso de la interfaz, etiquetado y gestión de alertas de calidad.
- **Seguimiento y feedback (Continuo)**: Revisión periódica de la salud de las ingestas, ajuste de reglas de observabilidad y expansión a nuevos departamentos.

Necesidades de formación del equipo
Capacitación técnica en la API de OpenMetadata para automatizaciones, formación en estándares de metadatos (JSON Schemas) y entrenamiento funcional para usuarios de negocio en la búsqueda y colaboración dentro de la plataforma.

Perfiles necesarios
- **Perfiles técnicos**: DevOps Engineer para el despliegue y mantenimiento, Data Architect para el diseño del modelo de metadatos y Data Engineer para la configuración de conectores e ingestas.
- **Personal externo recomendado**: Consultores expertos en Data Governance para los procesos de calidad y definición del glosario si la empresa carece de experiencia previa.
- **Otros**: Data Stewards (responsables de la veracidad de los metadatos) y Product Owners de datos.

Retorno de la inversión
- **Tiempos**: Se estima una mejora del 30-40% en la productividad de los analistas al reducir el tiempo de búsqueda y validación de datos en los primeros 6 meses.
- **Cómo medirlo, KPIs**: Tiempo medio de descubrimiento de datos (MTTD), porcentaje de activos de datos documentados, número de incidentes de calidad detectados proactivamente antes de llegar al reporte final, y reducción de tickets de soporte solicitando acceso o explicación de tablas.

Otros
OpenMetadata utiliza un enfoque basado en esquemas estándar que evita la dependencia del proveedor. Es compatible con el protocolo MCP (Model Context Protocol), lo que facilita que agentes de Inteligencia Artificial consuman y comprendan el contexto de los datos de la empresa de forma segura.

PREGUNTAS FRECUENTES

¿Qué es OpenMetadata y cuál es su función principal?

OpenMetadata es una plataforma de código abierto diseñada para la gestión unificada de metadatos bajo un estándar común. Su función principal es centralizar el descubrimiento de datos, el linaje, la observabilidad y el gobierno en una única interfaz, permitiendo que tanto equipos técnicos como de negocio colaboren y comprendan el ecosistema de datos de su organización de forma automatizada.

¿Para qué sirve el linaje de datos de extremo a extremo en esta herramienta?

Sirve para rastrear el flujo de la información desde su origen hasta su consumo final, permitiendo visualizar transformaciones incluso a nivel de columna individual. Es fundamental para realizar análisis de impacto, detectando qué dashboards o procesos se verán afectados si se modifica una estructura en el almacén de datos (Data Warehouse).

¿Cuánto cuesta utilizar OpenMetadata?

OpenMetadata ofrece una versión gratuita bajo licencia Open Source (Apache 2.0) que incluye todas las funcionalidades pero requiere que la organización gestione su propia infraestructura. Para empresas que prefieren un entorno gestionado, existe Collate, una versión SaaS con soporte profesional y funcionalidades avanzadas cuyo coste se establece bajo presupuesto personalizado.

¿Es OpenMetadata una solución de código abierto y puedo descargarla de GitHub?

Sí, es un proyecto plenamente Open Source distribuido bajo la licencia Apache 2.0. El código fuente, la documentación y los binarios necesarios para su despliegue están disponibles públicamente en su repositorio oficial de GitHub, donde cuenta con una comunidad activa de desarrolladores.

¿Cumple con la normativa española y europea de protección de datos (RGPD)?

La plataforma facilita el cumplimiento del RGPD mediante funcionalidades específicas como el etiquetado automático de información personalmente identificable (PII) y la implementación de políticas de control de acceso basadas en roles (RBAC). Esto permite a los responsables de datos identificar y proteger la información sensible de forma centralizada.

¿Qué nivel de seguridad y privacidad ofrecen sus datos?

OpenMetadata utiliza esquemas JSON estandarizados para definir metadatos, garantizando que el usuario mantenga la propiedad total de su información sin bloqueos de proveedor. La seguridad se gestiona mediante integraciones con protocolos estándar como SSO y OAuth, además de permitir auditorías completas a través de un registro histórico de cambios (versionado de metadatos).

¿Qué requisitos técnicos son necesarios para su instalación?

El despliegue requiere un nivel técnico alto, con conocimientos sólidos en la orquestación de contenedores mediante Docker y Kubernetes (Helm Charts). Además, es necesaria la configuración de servicios subyacentes como Elasticsearch o OpenSearch para las búsquedas y bases de datos como MySQL o PostgreSQL para el almacenamiento de metadatos.

¿Con qué herramientas y plataformas se puede integrar?

Dispone de más de 84 conectores nativos para las principales tecnologías del mercado, incluyendo bases de datos y almacenes de datos (Snowflake, BigQuery, Redshift), herramientas de transformación como dbt, orquestadores como Airflow, y plataformas de BI como PowerBI, Tableau y Looker. También ofrece una API REST extensiva y soporte para el protocolo MCP para agentes de IA.

¿Ofrece capacidades de observabilidad y calidad de datos?

Sí, permite implementar pruebas de calidad de datos sin necesidad de escribir código (no-code) y realizar un perfilado automático de los activos para monitorizar su salud. Esto ayuda a garantizar que la información consumida por los analistas sea fiable y esté actualizada.

¿Es adecuado para una pequeña empresa?

Generalmente no se recomienda para pequeñas empresas con infraestructuras mínimas o una única fuente de datos, ya que la complejidad de despliegue y mantenimiento de los servicios necesarios en Kubernetes suele superar los beneficios obtenidos en entornos de baja escala.

CONTRATOS Y CONDICIONES

Principales recomendaciones
- Realizar una evaluación de impacto en la protección de datos (EIPD) si se activa la detección automática de datos personales (PII) mediante escaneo de fuentes de datos.
- Configurar estrictamente el control de acceso basado en roles (RBAC) para limitar quién puede visualizar metadatos sensibles (como nombres de tablas que revelen estrategias comerciales o datos protegidos).
- En el caso de despliegue on-premise (Apache 2.0), la empresa es la única responsable de la seguridad de la infraestructura y el cumplimiento del RGPD frente a terceros.
- Si se utiliza la versión Cloud (Collate), es imperativo firmar un Acuerdo de Encargado de Tratamiento (DPA) con el proveedor para regular el acceso a los metadatos.
- Desactivar o supervisar el uso de conectores que realicen perfiles de datos (Profiling) si estos incluyen muestras de datos reales de clientes en entornos de desarrollo o gobierno.

Privacidad y protección de datos
- Responsabilidades: La empresa española actúa como Responsable del Tratamiento. OpenMetadata/Collate actúa como Encargado del Tratamiento solo en su modalidad Cloud.
- Ubicación de los datos: En la versión Open Source, los datos residen donde la empresa decida (España/UE recomendado). En la versión Collate, debe verificarse la región del tenant para evitar transferencias fuera del Espacio Económico Europeo.
- Transferencia internacional: El uso de la versión Cloud podría implicar transferencias a EE.UU. dependiendo de la ubicación de los servidores de Collate Inc.; se requiere verificar la adhesión al Data Privacy Framework.
- Derechos ARCO: La plataforma permite la trazabilidad de datos personales a través del linaje y el glosario, facilitando la identificación de dónde se almacenan datos de un interesado para cumplir con derechos de supresión o acceso.

Propiedad intelectual
- Propiedad de datos: Los metadatos integrados y generados pertenecen íntegramente a la empresa usuaria.
- Propiedad del resultado: El software Open Source está sujeto a la licencia Apache 2.0, que permite el uso comercial y la modificación, pero no otorga propiedad sobre el código base del motor.
- Licencias: Apache License 2.0 para la versión comunitaria; Licencia comercial propietaria para las funciones avanzadas de Collate (como AI Studio).

Usos y prohibiciones
- Usos admitidos: Gestión de gobernanza, auditoría de calidad de datos, automatización de catálogos y cumplimiento normativo interno.
- Usos prohibidos: No se debe utilizar para almacenar datos personales reales en los campos de descripción o etiquetas (solo deben contener metadatos), salvo que el campo esté cifrado y debidamente auditado.

Seguridad y certificaciones
- Seguridad: La herramienta soporta autenticación mediante SSO (Single Sign-On) y protocolos OAuth/OIDC (Google, Okta, Azure AD), recomendados para el cumplimiento del Esquema Nacional de Seguridad (ENS) en su nivel básico/medio.
- Certificaciones: La versión Collate (Cloud) busca habitualmente certificaciones SOC2 Type II; en la versión Open Source, la certificación de la infraestructura depende exclusivamente de la empresa que la hospeda.

Otros
- Es vital diferenciar entre el dato (contenido en la base de datos) y el metadato (información sobre el dato). OpenMetadata gestiona metadatos, pero su función de "Data Profiling" y "Sample Data" puede extraer muestras de datos reales, lo que eleva el riesgo legal de bajo a medio.

Fuentes consultada:
- Contratos: https://www.getcollate.io/terms-of-service
- Certificaciones: https://docs.open-metadata.org/v1.5.x/deployment/security
- Condiciones: https://www.getcollate.io/privacy-policy
- Licencias: https://github.com/open-metadata/OpenMetadata/blob/main/LICENSE

Para más información y herramientas:

Explora look4.tools para descubrir las mejores soluciones tecnológicas del mercado.

Inicio Todas las herramientas Categorías