

The **database** your AI agents deserve

Lakebase is serverless Postgres for applications that scale

[Explore the product](#) [See demo](#)

Project Dashboard

Monitoring

Branch: production Default Compute: Primary Active Refresh

JAN 15, 2026 4:21:14 PM

- CPU usage: 0.04
- RAM usage: 1.67 GB

1 Branch Project Settings

TRUSTED BY DATA + AI TEAMS

BURBERRY FDA Heineken MICHELIN Mercedes-Benz Santander SIEMENS

THE DATABRICKS PLATFORM

Build and run apps, agents and AI on your data

Databricks

Plataforma de Data Intelligence basada en arquitectura Lakehouse diseñada para empresas que necesitan centralizar grandes volúmenes de datos. Es la solución ideal para equipos de ingeniería de datos, científicos de datos y analistas que buscan ejecutar cargas de trabajo de ETL, Machine Learning y análisis SQL en un entorno colaborativo unificado sobre nubes públicas como AWS, Azure o GCP, eliminando silos de datos y simplificando la gobernanza mediante Unity Catalog y Delta Lake.

[Visitar Sitio Oficial](#) [Preguntar a ChatGPT](#) [Preguntar a Claude](#) [Preguntar a Grok](#)

Contenido del Dossier

- [Información de la Herramienta](#)
- [Consejos de Implantación](#)
- [Preguntas Frecuentes](#)
- [Contratos y Condiciones](#)

INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es

Databricks es una plataforma de "Data Intelligence" basada en una arquitectura unificada denominada Lakehouse, que combina la flexibilidad de un data lake con la gestión y rendimiento de un data warehouse. Está diseñada para empresas que necesitan centralizar grandes volúmenes de datos para ingeniería de datos, ciencia de datos, aprendizaje automático (ML) y análisis SQL en un único entorno colaborativo. En el ámbito profesional, se dirige a organizaciones con una mentalidad data-driven que operan en nubes públicas (AWS, Azure o GCP) y buscan eliminar los silos de datos entre sus equipos de ingeniería y analítica.

Principal ventaja profesional

La capacidad de ejecutar todas las cargas de trabajo de datos (ETL, BI y ML) sobre una única fuente de verdad (Delta Lake) con un motor de alto rendimiento (Photon), lo que simplifica radicalmente la arquitectura tecnológica y reduce los costes operativos de mantenimiento de múltiples sistemas.

Para quién no es

No es una herramienta adecuada para pequeñas empresas o profesionales independientes que solo requieren análisis de datos simples a partir de hojas de cálculo o bases de datos pequeñas. Tampoco es ideal para organizaciones que no tengan presencia en la nube o que carezcan de personal técnico para gestionar clusters de computación.

Funcionalidades clave

- **Unity Catalog:** Sistema de gobernanza unificada para datos y activos de IA en múltiples nubes.
- **Delta Lake:** Capa de almacenamiento de código abierto que aporta fiabilidad (transacciones ACID) a los data lakes.
- **Photon:** Motor de consulta vectorizado de última generación compatible con Apache Spark que acelera drásticamente el rendimiento.
- **Mosaic AI:** Herramientas integradas para el ciclo de vida completo de modelos de Machine Learning (MLflow) y despliegue de modelos generativos.
- **Databricks SQL:** Entorno optimizado para analistas de datos que permite ejecutar consultas SQL directamente sobre el data lake con rendimiento de data warehouse.
- **Notebooks Colaborativos:** Entorno compartido compatible con Python, SQL, R y Scala.

Precios

El modelo de precios es de pago por uso basado en **DBUs (Databricks Units)**, una unidad de medida de capacidad de procesamiento por segundo. El coste total se compone de la tarifa de Databricks más el coste de la infraestructura de la nube (máquinas virtuales, almacenamiento y red).

- **Versión gratuita:** Community Edition (limitada a un cluster pequeño de 15GB, uso educativo/prototipado) y prueba gratuita de 14 días de la plataforma completa.
- **Rango de precios:** Variable según el tipo de carga de trabajo (desde ~0.07\$ por DBU para Model Serving hasta ~0.70\$ por DBU para Serverless SQL).
- **Versiones de pago:**
 - **Premium:** Incluye Unity Catalog, seguridad avanzada, RBAC y auditoría.
 - **Enterprise:** Añade cumplimiento regulatorio (HIPAA, PCI-DSS, FedRAMP), claves de cifrado gestionadas por el cliente y controles de seguridad adicionales.
 - **Contratos de uso comprometido:** Descuentos significativos (hasta el 37%) mediante pagos por adelantado o compromisos anuales.

Perfil del usuario

- Empresas medianas y grandes en sectores como finanzas, retail, farma y tecnología que gestionan TB o PB de datos.
- **Perfiles profesionales:**
 - Ingenieros de Datos (ETL y pipelines).
 - Científicos de Datos y especialistas en ML.
 - Analistas de BI y Analytics Engineers.
 - Arquitectos de Soluciones Cloud.
 - CDO (Chief Data Officers) que buscan gobernanza centralizada.

Nivel técnico requerido

- **Para su uso:** Medio-Alto. Requiere conocimientos de SQL, Python, R o Scala, además de fundamentos de procesamiento distribuido (Spark).
- **Para instalación/configuración:** Alto. Exige conocimientos de administración de nubes (IAM, VPCs, almacenamiento de objetos) y seguridad de red.
- **Soporte:** Necesita soporte del departamento de IT/Cloud para la gestión de costes y seguridad.

Ejemplos de uso profesional

- Consolidación de múltiples bases de datos fragmentadas en un único repositorio de datos gobernado.
- Creación de pipelines de datos en tiempo real para detección de fraude financiero.
- Entrenamiento y despliegue masivo de modelos de recomendación personalizada en e-commerce.
- Generación de cuadros de mando ejecutivos que consultan directamente datos en bruto sin necesidad de moverlos a un almacén externo.

Uso y distribución

- **Versión web:** Acceso principal a través de un espacio de trabajo (Workspace) en el navegador.
- **Versión móvil:** Aplicación para monitoreo de jobs y alertas disponible en iOS/Android.
- **CLI:** Databricks CLI para automatización y gestión de recursos mediante comandos.
- **IDE:** Integraciones con VS Code y PyCharm (Databricks Connect).

Open Source

La plataforma se basa y contribuye a proyectos de código abierto fundamentales: Apache Spark, Delta Lake, MLflow y Koalas.

Integraciones

- **Facilidad de integración:** High Code / Low Code.
- **API propia:** REST API completa para automatizar la creación de clusters, ejecución de jobs y gestión de usuarios.
- **Integraciones nativas:** Conectores con herramientas de BI (Power BI, Tableau, Looker), herramientas de ingestión (Fivetran, Airbyte) y orquestadores (dbt, Airflow).

Notas finales

Información legal, licencias y contratos

- Se ofrece bajo un modelo de suscripción SaaS en nubes de terceros. El contrato suele ser un "Master Service Agreement" que especifica que el cliente mantiene la propiedad total de sus datos almacenados en su propia cuenta de la nube (S3, ADLS o GCS).

Otros

- Es importante monitorizar la "auto-terminación" de clusters para evitar costes inesperados por recursos inactivos.

Para más información:

- [Sitio web oficial](#)
- [Precios detallados](#)
- [Calculadora de costes](#)
- [Github oficial](#)
- [LinkedIn](#)

CONSEJOS DE IMPLANTACIÓN

Aplicación profesional

Databricks se posiciona como una plataforma de alto nivel para empresas que gestionan infraestructuras de datos complejas. Es ideal para organizaciones que manejan volúmenes desde Terabytes hasta Petabytes de información y requieren unificar procesos de Business Intelligence con Inteligencia Artificial. El presupuesto debe contemplar no solo el coste de las Unidades de Databricks (DBUs), sino también los costes de computación de proveedores como AWS, Azure o GCP. Es una herramienta clave en sectores regulados (Finanzas, Salud) por sus capacidades de cumplimiento (HIPAA, SOC2, PCI-DSS) y para empresas tecnológicas que escalan modelos de Machine Learning.

Madurez digital requerida

- Los usuarios deben dominar lenguajes de programación como SQL y Python, además de entender los fundamentos de la computación distribuida y el ecosistema de Apache Spark. No es apto para equipos que dependen exclusivamente de herramientas No-Code.
- La organización debe operar ya en una nube pública y tener establecidos procesos de gobernanza de datos. Se requiere una estructura departamental coordinada entre equipos de Ingeniería de Datos e IT para la gestión de infraestructura y seguridad.

Plan orientativo de implantación

Pasos necesarios y estimaciones

- Tiempos estimados de despliegue: Entre 3 y 6 meses para un despliegue operativo completo en producción, dependiendo de la complejidad de las fuentes de datos.
- Evaluación inicial (2-3 semanas): Auditoría de fuentes de datos, definición de casos de uso prioritarios y configuración del presupuesto de Cloud.
- Implantación inicial (4-6 semanas): Configuración del Workspace, establecimiento de la red (VPC Peering/Private Link) e implementación de Unity Catalog para la gobernanza.
- Prueba de concepto / Piloto (4-8 semanas): Migración de un pipeline ETL real o entrenamiento de un modelo de ML específico para validar rendimiento (Photon) y costes.
- Formación y despliegue (Continuo): Capacitación del equipo técnico y apertura del entorno a analistas de negocio mediante Databricks SQL.
- Seguimiento y Feedback: Monitorización semanal de costes de clusters y optimización de políticas de auto-terminación.

Necesidades de formación del equipo

Es indispensable formar al equipo en el uso de Delta Lake para la fiabilidad de las transacciones y en Unity Catalog para la gestión de permisos. Se recomienda certificación oficial en Databricks Data Engineer o Data Scientist para maximizar la eficiencia del código y evitar el desperdicio de créditos DBU por consultas mal optimizadas.

Perfiles necesarios

- Administrador de Cloud (AWS/Azure/GCP) para la infraestructura subyacente.
- Ingeniero de Datos (Data Engineer) para la creación de pipelines y gestión de clusters.
- Arquitecto de Datos para el diseño del modelo Lakehouse.
- Consultoría externa especializada para la configuración inicial de seguridad y optimización de costes si el equipo interno no tiene experiencia previa en Spark.

Retorno de la inversión

- El retorno se observa habitualmente entre los 6 y 12 meses tras la eliminación de silos de datos y la reducción de procesos manuales de movimiento de datos entre Data Lakes y Warehouses.
- KPIs: Reducción del tiempo de ejecución de procesos ETL (benchmarks de Photon muestran hasta 8x de mejora), disminución del tiempo de puesta en producción de modelos de ML y ahorro en costes de mantenimiento de infraestructuras duplicadas.

Otros

Es crítico implementar políticas de limitación de presupuesto (Budget Policies) y etiquetado de clusters (Tagging) desde el primer día para imputar costes a departamentos específicos y evitar sorpresas en la facturación mensual por clusters dejados en ejecución. El uso de Serverless SQL es recomendable para reducir la carga de gestión de infraestructura en tareas analíticas.

PREGUNTAS FRECUENTES

¿Qué es Databricks y en qué se diferencia de un data warehouse tradicional?

Databricks es una plataforma de inteligencia de datos basada en una arquitectura unificada denominada Lakehouse. A diferencia de los data warehouses tradicionales que suelen ser rígidos y costosos para datos no estructurados, el Lakehouse combina la flexibilidad y el bajo coste de un data lake con las capacidades de gestión, rendimiento y transaccionalidad ACID propias de un almacén de datos, permitiendo ejecutar ingeniería de datos, IA y analítica SQL en un mismo entorno.

¿Para qué sirve exactamente en un entorno profesional?

Sirve para centralizar y unificar el ciclo de vida completo del dato. Permite realizar procesos de extracción y transformación (ETL), ejecutar análisis de inteligencia de negocio (BI), desarrollar modelos de aprendizaje automático (ML) y gestionar la gobernanza de datos a gran escala, eliminando los silos de información entre equipos técnicos y de negocio.

¿Cómo funciona su modelo de costes y cuánto cuesta?

El modelo se basa en el pago por uso mediante Databricks Units (DBUs), que miden la capacidad de procesamiento por segundo. El coste final es la suma de la tarifa de Databricks (que varía según la carga de trabajo, desde 0.07\$ a 0.70\$ por DBU aproximadamente) y el coste de la infraestructura de computación y almacenamiento del proveedor de nube (AWS, Azure o GCP) utilizado.

¿Existe una versión gratuita o para pruebas?

Sí, existe la 'Community Edition', una versión gratuita con capacidades limitadas a un cluster de 15GB diseñada para fines educativos y prototipado sencillo. Además, la plataforma ofrece una prueba gratuita de 14 días para evaluar todas las funcionalidades de las versiones empresariales en un entorno de producción.

¿Es Databricks una tecnología de código abierto (Open Source)?

La plataforma como servicio (SaaS) es propietaria, pero está construida íntegramente sobre estándares de código abierto creados o liderados por sus fundadores, como Apache Spark, Delta Lake y MLflow. Esto permite evitar el bloqueo por proveedor (vendor lock-in) y facilita la portabilidad de las cargas de trabajo.

¿Se puede descargar y gestionar desde GitHub?

No se descarga la plataforma completa, ya que es un servicio gestionado en la nube. Sin embargo, Databricks mantiene numerosos repositorios oficiales en GitHub que incluyen el código de sus proyectos abiertos (Delta, MLflow), conectores (Databricks Connect), herramientas de línea de comandos (CLI) y plantillas de implementación.

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

Sí, Databricks es compatible con el RGPD (GDPR). Su arquitectura permite que los datos permanezcan en la cuenta de nube del cliente (dentro de regiones geográficas específicas como España o la UE), y sus versiones Enterprise cumplen con certificaciones de seguridad adicionales como ISO 27001, SOC 2, HIPAA y PCI-DSS.

¿Cómo garantiza la seguridad y la privacidad de los datos?

La plataforma utiliza Unity Catalog para ofrecer una gobernanza centralizada con control de acceso basado en roles (RBAC) y auditorías detalladas. La privacidad se refuerza mediante el cifrado de datos en reposo y en tránsito, con la opción en planes superiores de gestionar las propias claves de cifrado (Customer-Managed Keys).

¿Qué nivel técnico se requiere para su implementación y uso?

Para la configuración inicial se requiere un nivel alto de conocimientos en administración de infraestructura cloud (redes, identidad y almacenamiento). Para el uso diario, los profesionales deben tener un nivel medio-alto en lenguajes como SQL, Python, R o Scala, además de comprender los fundamentos del procesamiento de datos distribuidos.

¿Es una tecnología segura frente a brechas de datos?

Es una tecnología diseñada para entornos empresariales críticos. Incluye aislamiento de red mediante VPC/VNet, autenticación multifactor y controles de seguridad perimetral. No obstante, la seguridad final depende de la correcta configuración de las políticas de acceso y la monitorización activa de los clusters por parte del equipo de IT de la organización.

CONTRATOS Y CONDICIONES

Informe técnico descriptivo: Databricks

Principales recomendaciones

- **Firma del DPA:** Es imprescindible que la empresa española firme el Adenda de Procesamiento de Datos (DPA) de Databricks para establecer las garantías exigidas por el RGPD.
- **Configuración de Residencia:** Al ser una herramienta que se despliega sobre nubes públicas (AWS, Azure o GCP), se debe configurar específicamente el despliegue en regiones dentro de la Unión Europea (ej. Madrid, Bélgica, Frankfurt) para minimizar riesgos de transferencias internacionales.
- **Uso de Unity Catalog:** Activar obligatoriamente esta función para centralizar la gobernanza, el control de accesos y la trazabilidad (lineage) de los datos, facilitando el cumplimiento de auditorías.
- **Anonimización en origen:** Evitar la ingesta de datos personales directos en el "Data Lake" si no es estrictamente necesario para el procesamiento, aplicando técnicas de seudonimización antes de la carga.

Ley de Inteligencia Artificial (AI Act)

- **Clasificación del sistema:** Dependiendo del uso (ej. scoring crediticio o selección de personal), los modelos desarrollados en Databricks pueden clasificarse como de "Alto Riesgo".
- **Gobernanza de modelos:** El uso de MLflow y Unity Catalog para IA permite cumplir con los requisitos de documentación técnica, registro de actividad (logs) y transparencia que exige la normativa.
- **Calidad de datos:** La plataforma facilita el cumplimiento del requisito de "datos de entrenamiento representativos y libres de sesgos" mediante herramientas de análisis y perfilado de datos integradas.

Privacidad y protección de datos

- **Responsabilidades:** La empresa española actúa como "Responsable del Tratamiento" y Databricks como "Encargado del Tratamiento".
- **Ubicación de los datos:** Los datos se almacenan en la cuenta de almacenamiento del cliente (S3, Azure Blob, GCS). La soberanía reside en la configuración de la nube elegida por la empresa.
- **Transferencia internacional:** Databricks utiliza Cláusulas Contractuales Tipo (SCCs) y está adherido al "Marco de Privacidad de Datos UE-EE. UU." para legitimar transferencias de metadatos o soporte técnico a EE. UU.
- **Derechos ARCO:** La plataforma ofrece herramientas para la ejecución técnica de derechos (borrado o acceso) mediante comandos SQL y APIs, permitiendo localizar y eliminar datos de sujetos específicos en tablas Delta.

Propiedad intelectual

- **Propiedad de datos:** La empresa cliente mantiene la propiedad total y exclusiva de todos los datos introducidos en la plataforma.
- **Propiedad del resultado:** El código desarrollado en los Notebooks (Python, SQL, etc.) y los modelos de IA resultantes pertenecen a la empresa cliente, salvo pacto en contrario basado en librerías de terceros con licencias específicas.

Usos y prohibiciones

- **Usos admitidos:** Análisis predictivo, ingeniería de datos, inteligencia de negocio (BI), entrenamiento de modelos de lenguaje (LLMs) y almacenamiento masivo gobernado.
- **Usos prohibidos:** El cliente no debe utilizar la plataforma para actividades ilícitas, vulnerar derechos de propiedad intelectual de terceros o procesar categorías especiales de datos (religión, salud avanzada, etc.) sin las medidas de seguridad "Enterprise" adicionales.

Seguridad y certificaciones

- **Seguridad:** Cifrado en reposo (AES-256) y en tránsito (TLS 1.2+). Soporte para claves gestionadas por el cliente (Customer Managed Keys) en versiones superiores.
- **Certificaciones:** Cumple con ISO 27001, ISO 27017, ISO 27018, SOC 2 Type II y Esquema Nacional de Seguridad (ENS) en nivel Alto (según el proveedor cloud subyacente).

Otros

- **Seguridad de Red:** Se recomienda el uso de "Private Link" para que el tráfico de datos no circule por la internet pública, conectando directamente la red de la empresa con el plano de control de Databricks.

Fuentes consultadas:

- [Contratos: Master Cloud Services Agreement](#)
- [Certificaciones: Trust Center](#)
- [Condiciones: Data Processing Addendum \(DPA\)](#)
- [Licencias: Open Source \(Apache/Delta\)](#)

Para más información y herramientas:

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

[Inicio](#) [Todas las herramientas](#) [Categorías](#)

Este documento ofrece recomendaciones generadas mediante análisis humano y sistemas de IA automatizados. La información tiene carácter meramente informativo y no constituye asesoramiento legal, profesional ni garantía de resultados. Las marcas, logotipos y nombres comerciales pertenecen a sus respectivos propietarios y se utilizan únicamente con fines identificativos.