

Introducción a Databricks y el ecosistema Spark

Databricks
Onboarding



IMMUNE
TECHNOLOGY INSTITUTE



¿Qué es Databricks?

Databricks es una plataforma de análisis de datos basada en la nube diseñada para simplificar el uso de Apache Spark.

Permite a los científicos de datos, ingenieros de datos y analistas colaborar en proyectos complejos de datos y machine learning (ML), ofreciendo una interfaz intuitiva y herramientas avanzadas para el procesamiento y análisis de grandes volúmenes de datos.

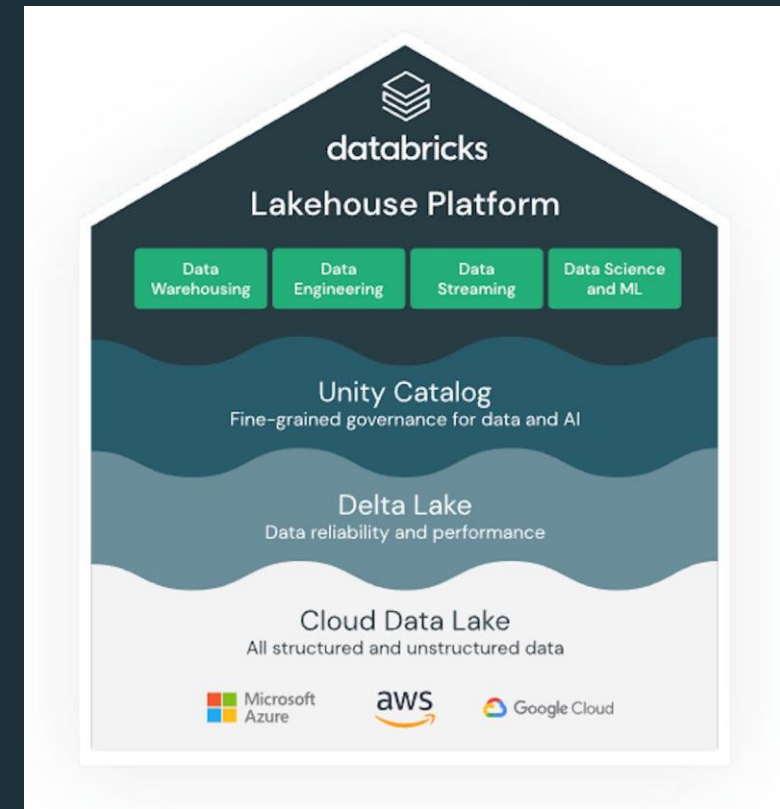


Databricks – Lakehouse

Lakehouse es una arquitectura moderna que integra la flexibilidad de los data lakes con las capacidades analíticas de los data warehouses, utilizando Delta Lake para garantizar el rendimiento y fiabilidad, y gestionar metadatos.

Facilita el manejo y análisis de grandes volúmenes de datos estructurados y no estructurados en una plataforma unificada.

Implementa rigurosos controles de gobernanza y seguridad, proporcionando una herramienta única para el gobierno de datos



Databricks vs. Entornos de Spark Tradicionales

Databricks se distingue de los entornos de Spark tradicionales al proporcionar una plataforma unificada que integra todas las herramientas necesarias para el procesamiento y análisis de datos y el desarrollo de modelos de ML.

A diferencia de trabajar con Apache Spark en un cluster de Hadoop o un entorno Spark on-premise, Databricks simplifica la configuración, gestión y escalabilidad de los recursos, permitiendo a los usuarios enfocarse directamente en los datos y el desarrollo de procesos.



Key Features

- **Integración nativa con la nube:** Facilita el acceso y análisis de datos almacenados en la nube.
- **Gestión automática de clusters:** Reduce el tiempo y la complejidad de configurar y mantener clusters de Spark.
- **Ambiente colaborativo:** Permite a los equipos trabajar juntos en notebooks compartidos con soporte para SQL, Python, R, y Scala.
- **Optimización de rendimiento:** Mejoras en la ejecución de Spark para procesar datos más rápido (Databricks Runtime, Autoscaling, Photon, Catalyst optimized)
- **Seguridad avanzada:** Incluye características de seguridad de nivel empresarial para proteger los datos y cumplir con regulaciones (Data Encryption, RBAC y MFA, Integración con ID corporativa (Azure AD, etc...), VNets)
- **Data Governance (Unity Catalog):** es un catálogo de gobernanza centralizado para la gestión de metadatos y usuarios, ofreciendo un modelo de permisos unificado a los datos.
- **Delta Share:** Permite el intercambio de datos entre organizaciones utilizando un protocolo estándar



Unity Catalog

Unity Catalog es una solución de gobierno unificado de datos en Databricks.

Se construye de forma nativa en la plataforma y proporciona una gestión centralizada de metadatos y usuarios, lo que permite un acceso y control consistentes a los datos.



Unity Catalog – Características

Gestión Centralizada de Metadatos: Unifica la administración de metadatos en todas las áreas de trabajo, lo que permite la colaboración y garantiza la coherencia en los controles de acceso y los datos.

Controles de Acceso Unificados: Centraliza los controles de acceso para archivos, tablas y vistas, utilizando vistas dinámicas para controles de acceso detallados, lo que permite restringir el acceso a filas y columnas según las autorizaciones de usuarios y grupos.

Interoperabilidad y Flexibilidad: Unity Catalog admite integraciones nativas con una variedad de herramientas y plataformas, lo que permite a los equipos de datos trabajar con las herramientas que elijan sin problemas de gobernanza o rendimiento.

Interfaz de Gobierno: Ofrece una interfaz de usuario intuitiva y capacidades de línea de comandos para administrar fácilmente el acceso a los datos y las políticas de seguridad en la nube.



Delta Share

Delta Share de Databricks es una innovadora función que revoluciona el intercambio de datos en tiempo real entre organizaciones, permitiendo compartir datos seguros y actualizados sin moverlos físicamente.

Esta herramienta elimina barreras de intercambio, facilita el cumplimiento normativo, mejora la colaboración, y abre caminos para obtener insights valiosos.

Integrado nativamente en Databricks, Delta Share apoya eficazmente el análisis de datos y proyectos de machine learning, simplificando la gestión de datos entre plataformas.



Delta Share – Características

Interoperabilidad y Estándares Abiertos: Delta Sharing utiliza un protocolo abierto para el intercambio de datos con cualquier plataforma que admita este estándar, asegurando la interoperabilidad entre diferentes sistemas y herramientas de análisis de datos.

Seguridad y Control de Acceso: Ofrece mecanismos robustos de seguridad y control de acceso, permitiendo a los propietarios de los datos definir quién puede acceder a sus datos y bajo qué condiciones. Esto incluye la gestión de permisos a nivel de datos, garantizando que solo los usuarios autorizados puedan acceder a la información compartida.

Eficiencia en la Transmisión de Datos: Optimiza la transferencia de datos a través de la red, reduciendo el tiempo y los recursos necesarios para compartir grandes volúmenes de datos. Utiliza técnicas como la compresión de datos y la transmisión de deltas (solo los cambios desde la última sincronización) para mejorar la eficiencia.



Delta Share – Características

Facilidad de Uso: Delta Sharing está diseñado para ser fácil de usar, permitiendo a las organizaciones compartir y acceder a datos compartidos con pocas configuraciones. Se integra de manera transparente con Databricks y otras plataformas de análisis de datos, simplificando el proceso de intercambio de datos.

Soporte para Datos en Tiempo Real: Permite el intercambio de datos en tiempo real, lo que es crucial para aplicaciones que dependen de la actualización constante de información, como el análisis en tiempo real, la inteligencia de negocios, y el aprendizaje automático. Esto facilita la colaboración en escenarios donde el tiempo es crítico.

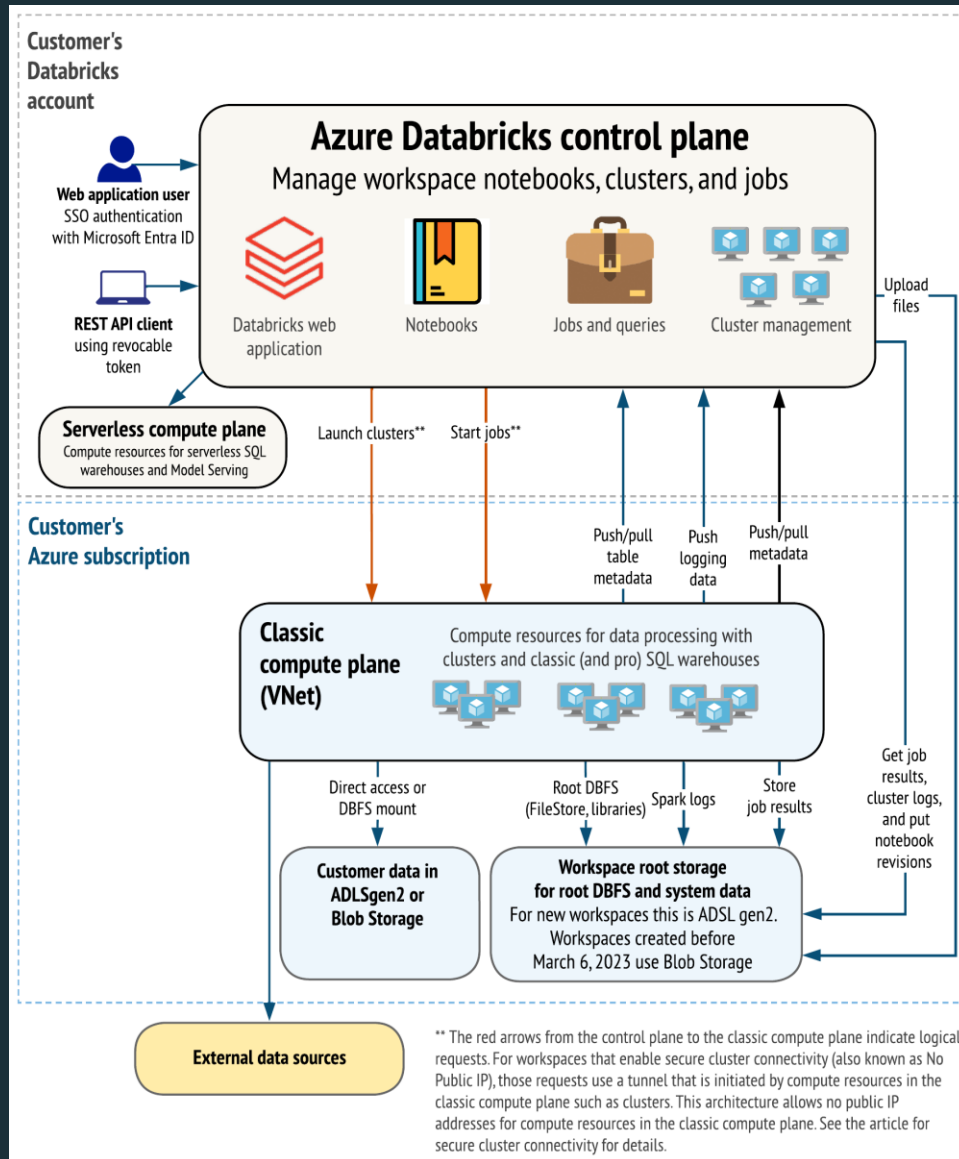


Arquitectura

- La arquitectura de Databricks se fundamenta en la separación de funcionalidades entre 2 planos o capas principales:
 - **Plano de control:** Databricks gestiona servicios de backend y almacenan y encriptan las configuraciones del espacio de trabajo, notebooks, librerías de código, etc. Este plano facilita la administración de los recursos y la colaboración entre equipos Data Scientists, Data analysts, y Data Engineers.
 - **Plano de cómputo:** Aquí se procesan los datos. Los recursos de cómputo se crean en la cuenta de nube del usuario y son gestionados dentro de una red virtual propia de la organización a la que pertenece usuario.
- Los datos se almacenan en el llamado **plano de datos**, que son los sistemas de almacenamiento propiedad de la organización a la que pertenecen los usuarios



Componentes



Pago por uso – Databricks Units (DBUs)

Al usar Databricks, generalmente hay dos conceptos principales que se pagan:

Databricks Units (DBUs): Un DBU es una unidad de potencia de procesamiento en la plataforma de Databricks, utilizada como medida y para fines de facturación.

Esta unidad está influenciada por el uso de recursos de cómputo y la cantidad de datos procesados. Por lo tanto, el coste aumenta con el tamaño del clúster y la configuración específica, independientemente de si se utiliza un solo nodo o un clúster de muchos nodos.

Los planes de precios varían desde opciones **estándar** hasta planes **premium** y **empresariales**, con cada plan ofreciendo diferentes características y tipos de cargas de trabajo disponibles.



Pago por uso – Cloud Infrastructure

Costes de la plataforma de nube: Estos son los costos de los recursos de cómputo (como instancias VM) proporcionados por el proveedor de servicios en la nube (AWS, GCP, Azure) que se utilizan en conjunto con Databricks.

Estos costes son adicionales a los costos de DBU y se facturan según el uso, normalmente a una tarifa por horas.



Pago por uso – Tabla precios VM Azure – Serie Dsv3

Instancia	vCPU	RAM	Número de DBU	Precio de la DBU	Precio total de Pago por uso	Precio total de la máquina virtual reservada de 1 año (% de ahorro)	Precio total de la máquina virtual reservada de 3 años (% de ahorro)	Precio total de las instancias de acceso puntual (ahorro en porcentaje)
D4s v3	4	16,00 GiB	0,75	€0,382/hora	€0,605/hora	€0,524/hora ~13 % de ahorro	€0,480/hora ~21 % de ahorro	€0,405/hora ~33 % de ahorro
D8s v3	8	32,00 GiB	1,50	€0,764/hora	€1,209/hora	€1,048/hora ~13 % de ahorro	€0,960/hora ~21 % de ahorro	€0,809/hora ~33 % de ahorro
D16s v3	16	64,00 GiB	3,00	€1,528/hora	€2,418/hora	€2,095/hora ~13 % de ahorro	€1,918/hora ~21 % de ahorro	€1,617/hora ~33 % de ahorro
D32s v3	32	128,00 GiB	6,00	€3,056/hora	€4,835/hora	€4,189/hora ~13 % de ahorro	€3,836/hora ~21 % de ahorro	€3,234/hora ~33 % de ahorro
D64s v3	64	256,00 GiB	12,00	€6,112/hora	€9,669/hora	€8,378/hora ~13 % de ahorro	€7,672/hora ~21 % de ahorro	€6,468/hora ~33 % de ahorro



Databricks UI

The screenshot displays the Databricks user interface. At the top, there's a header with 'Microsoft Azure', the 'databricks' logo, a search bar, and a 'CTRL + P' shortcut. A left sidebar contains navigation links for 'New', 'Workspace', 'Recents', 'Catalog', 'Workflows', 'Compute', 'SQL', 'SQL Editor', 'Queries', 'Dashboards', 'Alerts', 'Query History', 'SQL Warehouses', 'Data Engineering', 'Job Runs', 'Data Ingestion', 'Delta Live Tables', 'Machine Learning', 'Experiments', 'Features', 'Models', and 'Serving'. The main area is divided into two panels. The left panel, titled 'Workspace', shows a tree view with 'Home', 'Workspace' (containing 'Shared' and 'Users'), 'Repos' (containing 'anrau232@outlook.com'), and 'Favorites'. The right panel, titled 'Repos', shows the 'databricks_workshop' repository with a table of its contents:

Name	Type
code labs	Folder
datasets	Folder
slide deck	Folder
README.md	File

La interfaz de usuario de Databricks está diseñada para ser accesible y fácil de usar, permitiendo a los usuarios navegar eficientemente a través de diversas funcionalidades como la gestión de clústers, el desarrollo y colaboración en notebooks, y la visualización de datos, todo desde un entorno centralizado



Databricks UI – Workspace

El Workspace de Databricks es el lugar donde los usuarios pueden crear, organizar y compartir notebooks y librerías. Es ideal para la colaboración en equipo, permitiendo la integración y análisis de datos en un entorno unificado

Workspace

> Home

Workspace

- Shared
- Users
 - anrau232@outlook.com
 - anrau232_outlook.com#ext#@anrau232outl...



Repos






- anrau232@outlook.com
 - databricks_workshop**
 - code labs
 - datasets
 - slide deck

Favorites

- Trash


Repos > anrau232@outlook.com >



databricks_workshop  main 

Name 	Type	Owner
 code labs	Folder	Jesus Arnau
 datasets	Folder	Jesus Arnau
 slide deck	Folder	Jesus Arnau
 README.md	File	Jesus Arnau







Databricks UI – Compute



Spark cluster 


Policy 
Unrestricted 




☒ Multi node ☐ Single node


Access mode  Single user access 
Single user  Arnau Villar, Jesus 



Performance



Databricks runtime version 
Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1) 



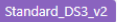

☒ Use Photon Acceleration 

Worker type  Min workers Max workers
Standard_DS3_v2 14 GB Memory, 4 Cores  2 8 ☐ Spot instances 

Driver type
Same as worker 14 GB Memory, 4 Cores 

☒ Enable autoscaling 
☒ Terminate after 120 minutes of inactivity 

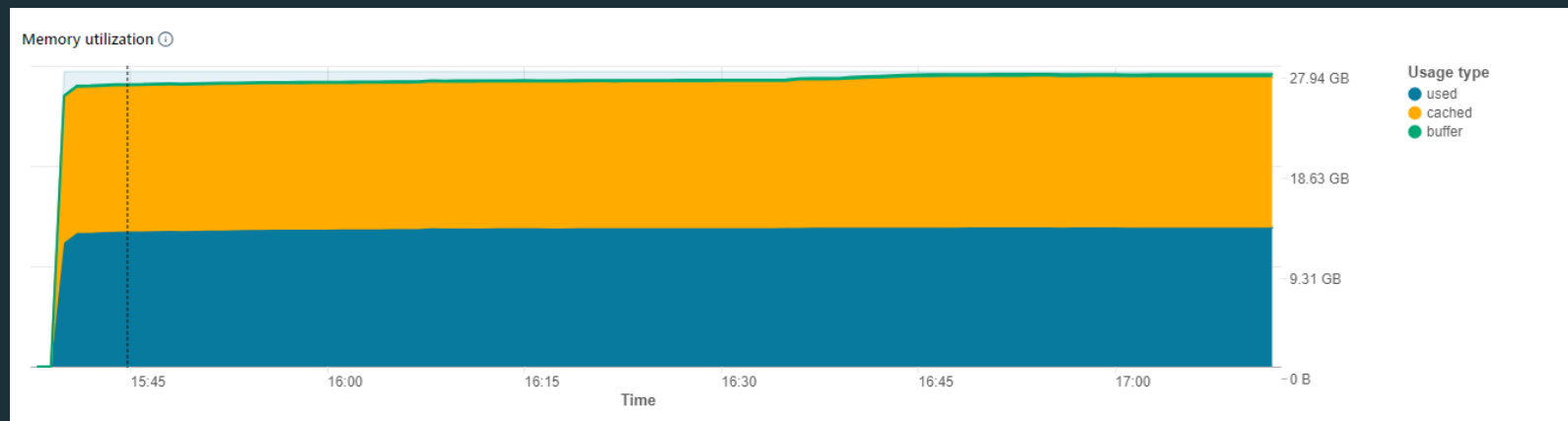
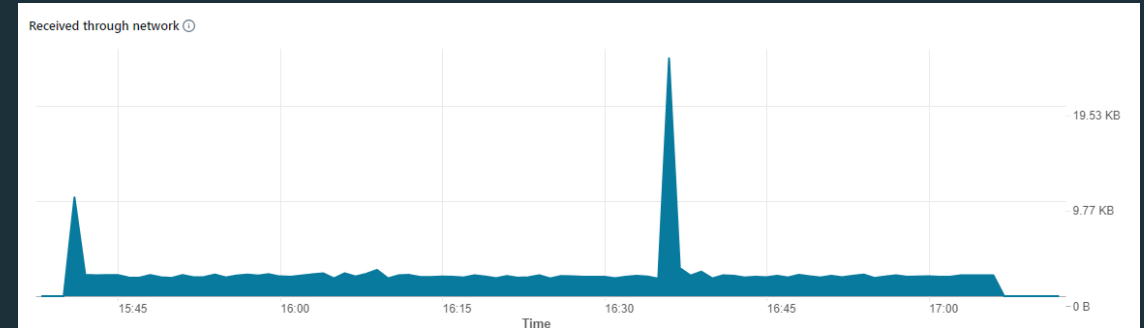
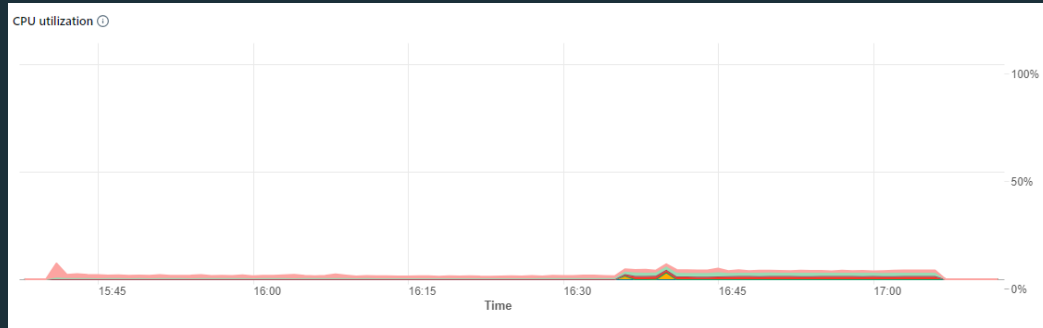
Tags 
Add tags
Key Value 
> Automatically added tags
▶ Advanced options

Summary
2-8 Workers 28-112 GB Memory 8-32 Cores
1 Driver 14 GB Memory, 4 Cores
Runtime 13.3.x-scala2.12
  


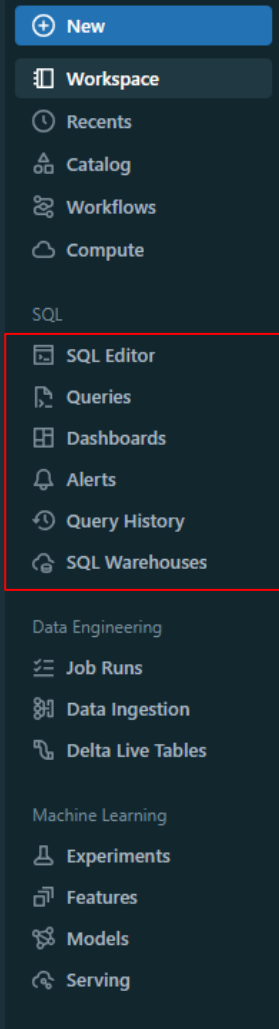
- La gestión de Clústers es fundamental en Databricks, permitiendo a los usuarios configurar, iniciar, y detener clusters de Spark según la demanda.
- Esto incluye opciones de autoescalado y monitoreo en tiempo real para optimizar el rendimiento y los costos



Databricks UI – Compute – Monitoring



Databricks UI – Casos de uso – Data Analytics

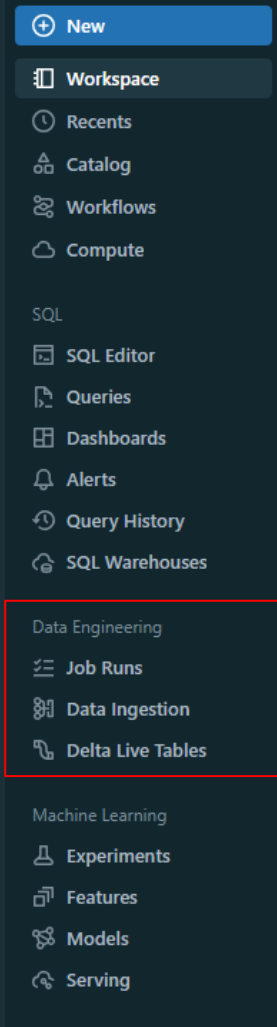


En el menú de navegación podremos acceder a los 3 casos de uso principales que se dan en la práctica totalidad de proyectos de Big Data:

- Data Analytics (SQL): Databricks permite realizar análisis SQL avanzados, facilitando la ejecución de consultas SQL para la generación de informes de BI, análisis y visualizaciones, extrayendo insights y KPI's desde un data lake.
- Ideal para Data Analysts que buscan realizar exploración de datos ad-hoc y desarrollar dashboards interactivos.



Databricks UI – Casos de uso – Data Engineering



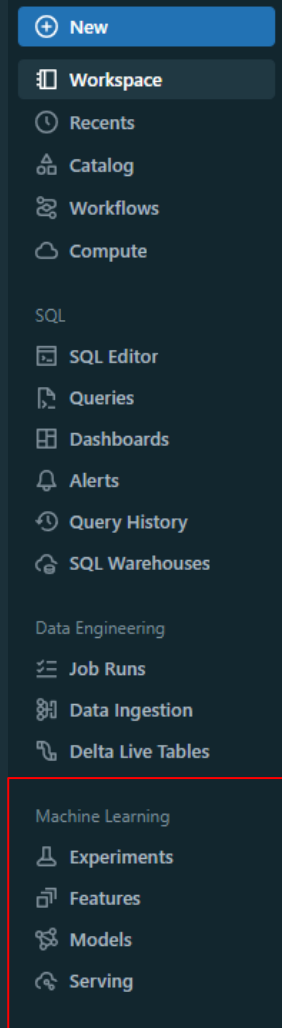
Databricks proporciona un entorno potente para la ingeniería de datos, donde se pueden construir pipelines de procesamiento de datos robustos y escalables.

Los Data Engineers pueden utilizar Databricks para automatizar y simplificar tareas complejas de ETL (Extracción, Transformación y Carga), permitiendo la limpieza, consolidación y preparación de datos procedentes de diversas fuentes.

Con la capacidad de manejar tanto procesos en streaming como batch, Databricks facilita el procesamiento de grandes conjuntos de datos (Big Data), integración con herramientas de streaming, y la gestión de lago de datos (Data Lake) mediante Delta Lake, que aporta funcionalidades transaccionales y esquema en evolución para un manejo más eficiente de los datos.



Databricks UI – Casos de uso – Data Engineering



La plataforma de Databricks ofrece un entorno integral para el desarrollo de Machine Learning, brindando a los Data Scientists e ingenieros de ML un conjunto de herramientas para construir, entrenar y desplegar modelos de aprendizaje automático con alta eficiencia y escalabilidad.

Databricks facilita la experimentación rápida con soporte para una amplia gama de frameworks de ML, incluyendo TensorFlow, PyTorch y scikit-learn, y se integra con MLflow para el seguimiento de experimentos, la gestión de modelos y el registro de artefactos.



Databricks Learning

Para profundizar en Databricks y sus capacidades, la plataforma ofrece una amplia gama de recursos, incluyendo documentación detallada, tutoriales, webinars, y un foro de la comunidad activo:

- Databricks Academy: [Academy Login | Databricks](#)
- Databricks Community: <https://community.databricks.com>
- Databricks Community Edition: [Login - Databricks Community Edition](#)



Lab 1: Databricks Walkthrough

