# Introducción a Databricks y el ecosistema Spark

Closing

Q&A

databricks

IMMUNE
TECHNOLOGY INSTITUTE

# Key Features

**Unificación de Data Warehousing y Data Lakes (Lakehouse)**: Unifica los silos de datos permitiendo almacenar y analizar datos estructurados y no estructurados en un solo lugar, optimizando así la gestión de datos y el análisis avanzado.

**Delta Lake:** Ofrece una capa de almacenamiento confiable que facilita la gestión de datos con operaciones ACID, garantizando la integridad y consistencia de los datos para análisis fiables.

**Data Analytics**: Proporciona una plataforma integrada que simplifica el análisis de datos a gran escala, permitiendo a los usuarios obtener insights rápidos y precisos sin la complejidad del manejo de infraestructuras.

# Key Features

**Fácil administración y escalado de la computación**: Automatiza la configuración, el escalado y la optimización de los recursos de computación, reduciendo la carga operativa y permitiendo a los equipos centrarse en el análisis de datos.

**Interoperabilidad y Ecosistema**: Se integra sin problemas con herramientas existentes y servicios en la nube, ofreciendo flexibilidad y acceso a un ecosistema extenso para potenciar los análisis de datos y la innovación.

# Databricks - Best Practices

- **Usar Job Clusters en lugar de All-Purpose Clusters para ejecutar procesos en producción**: Los Job Clusters son específicos para una tarea o conjunto de tareas y se apagan automáticamente una vez que el trabajo está completo, lo que ayuda a gestionar mejor los costos al evitar el uso de recursos innecesarios.
- **Optimizar el tamaño y el tipo de instancias según la carga de trabajo**: Seleccionar el tamaño y tipo de instancia adecuados para tu cluster puede tener un gran impacto en el rendimiento y el costo. Databricks ofrece una variedad de tipos de instancias optimizadas para diferentes cargas de trabajo, como computación intensiva o cargas de trabajo con uso intensivo de memoria.
- **Particionar y optimizar los datos almacenados en Delta Lake**: La partición de datos permite un acceso más rápido y eficiente a los datos al dividirlos en subconjuntos más pequeños basados en columnas clave. Optimizar las tablas de Delta Lake mejora el rendimiento de las consultas y reduce los costos de lectura y escritura.

# Databricks – Best Practices

- **Utilizar Databricks Runtime Optimizado**: Databricks ofrece runtimes optimizados que están preconfigurados con las mejores configuraciones para diferentes tipos de cargas de trabajo. Esto incluye mejoras de rendimiento para operaciones de Spark y ML, así como integraciones optimizadas para bibliotecas comunes.
- **Aplicar políticas de acceso y seguridad de datos**: Implementar políticas de control de acceso basado en roles (RBAC) para asegurar que solo los usuarios autorizados puedan acceder a los datos y ejecutar trabajos. Esto es crucial para mantener la seguridad y cumplir con las regulaciones de datos.

# Recursos Adicionales

- Databricks Blog: Databricks Blog
- Databricks Platform Release Notes: Databricks platform release notes | Databricks on AWS
- Databricks Academy: Academy Login | Databricks
- Databricks Community: https://community.databricks.com
- Databricks Community Edition: Login - Databricks Community Edition
- Codigo de los laboratorios y slides: jarnawer/databricks_workshop (github.com)

databricks

# Questions & Answers

IMMUNE
TECHNOLOGY INSTITUTE

databricks


# MUCHAS GRACIAS!!!

IMMUNE
TECHNOLOGY INSTITUTE