RHCECAM
CONSULTORÍA Y CAPACITACIÓN

CURSO DE DATABRICKS

Modalidad Live Stream

De Fundamentos hasta Producción · 20 Horas · Nivel Básico a Avanzado

8	20	Basic → Advanced	Labs incluidos
Módulos	Horas totales	Nivel	Práctica

Descripción del Curso

Este programa de 20 horas lleva al participante desde cero hasta dominar los componentes clave de la plataforma Databricks en entornos de producción. Combina sesiones teóricas, laboratorios prácticos en Notebooks y casos de uso reales con PySpark, Delta Lake, SQL, MLflow y Workflows.

Dirigido a:

- Data Engineers que buscan certificarse o migrar pipelines a Databricks
- Analistas de datos con conocimiento de SQL que quieren escalar a Big Data
- Científicos de datos o actuarios que necesitan llevar modelos a producción
- Arquitectos de soluciones que diseñan plataformas analíticas modernas

Requisitos previos:

- Conocimiento básico de Python (variables, funciones, loops)
- Familiaridad con SQL (SELECT, JOIN, GROUP BY)
- Conceptos básicos de datos estructurados y no estructurados

Distribución de Módulos

#	Módulos	Nivel	Horas
01	Fundamentos de Databricks y Lakehouse	Básico	2h 30m
02	Ambiente de Desarrollo: Notebooks y Clusters	Básico	1h 30m
03	Apache Spark con PySpark	Básico-Intermedio	3h 00m
04	Delta Lake: Gestión de Datos Confiable	Intermedio	2h 30m
05	Databricks SQL y Visualización	Intermedio	2h 00m
06	Pipelines de Datos con Delta Live Tables	Intermedio-Avanzado	3h 00m
07	MLflow y Machine Learning en Databricks	Avanzado	2h 30m
08	Producción, Seguridad y Optimización	Avanzado	3h 00m
	TOTAL		**20h 00m**

Contenido Detallado por Módulo

01	Fundamentos de Databricks y Lakehouse	2h 30m
	¿Qué es Databricks?	*Historia, casos de éxito, posicionamiento vs Spark standalone*
	Arquitectura Lakehouse	*Data Warehouse vs Data Lake vs Lakehouse, Delta Lake como cimiento*
	Workspace y Proveedores Cloud	*AWS, Azure, GCP — diferencias de setup y networking*
	Autenticación y Acceso	*Personal Access Tokens, Service Principals, SSO/SCIM*
	Navegación del Workspace	*Repos, DBFS, Unity Catalog overview, menú de recursos*
	Lab 01	*Crear workspace trial, explorar UI, conectar DBFS Explorer*

02	Ambiente de Desarrollo: Notebooks y Clusters	1h 30m
	Notebooks avanzados	*Magic commands %sql %md %sh, widgets, parámetros dinámicos*
	Tipos de Clusters	*All-purpose vs Job Clusters, fotones, autoscaling, spot instances*
	Configuración de Clusters	*Runtime versions, init scripts, Spark configs clave*
	Gestión de Librerías	*PyPI, Maven, wheel files, cluster-scoped vs notebook-scoped*
	Lab 02	*Crear cluster con autoscaling, instalar librerías, correr notebook con widgets*

03	Apache Spark con PySpark	3h 00m
	Arquitectura Spark	*Driver, Executors, Slots, Tasks, Stages, DAG Scheduler*
	DataFrames y Transformaciones	*Lazy evaluation, narrow vs wide transformations, Actions vs Transformations*
	SQL en Spark	*spark.sql(), createOrReplaceTempView, SparkSession config*
	Optimizaciones	*Particionado, broadcast joins, cache/persist, repartition vs coalesce*
	Spark UI & Tuning	*DAG Visualizer, diagnóstico de skew, AQE, cost-based optimizer*
	Lectura/Escritura Multi-formato	*CSV, JSON, Parquet, ORC, Avro — schemas, inferencia, opciones*
	Lab 03	*Pipeline ETL completo: ingest CSV → transformar con PySpark → escribir Parquet particionado*

04	Delta Lake: Gestión de Datos Confiable	2h 30m
	Fundamentos de Delta Lake	*Transaction Log, versioning, ACID en el Lakehouse*
	Operaciones DML	*INSERT, UPDATE, DELETE, MERGE — upserts a escala*
	Time Travel	*VERSION AS OF, TIMESTAMP AS OF, RESTORE TABLE, auditoría*
	Optimización	*OPTIMIZE, ZORDER, Auto Optimize, file compaction, particiones*
	Change Data Feed	*Habilitación, lectura de cambios, casos CDC streaming*
	Schema Enforcement y Evolution	*mergeSchema, overwriteSchema, column mapping*
	Lab 04	*Crear tabla Delta, ejecutar MERGE upsert, hacer Time Travel y restaurar versión*

05	Databricks SQL y Visualización	2h 00m
	Databricks SQL Workspace	*SQL Editor, Query History, Query Profiles*
	SQL Warehouses	*Serverless vs Classic, T-shirt sizes, Auto Stop, concurrencia*
	Objetos del Catálogo	*Schemas, Tables (Managed vs External), Views, Functions en Unity Catalog*
	Dashboards y Alertas	*Lakeview Dashboards, visualizaciones nativas, alertas por threshold*
	Conectores BI	*Power BI, Tableau, Looker via Partner Connect, JDBC/ODBC*
	Lab 05	*Crear SQL Warehouse, escribir queries analíticas sobre Delta, publicar dashboard*

06	Pipelines de Datos con Delta Live Tables	3h 00m
	Introducción a DLT	*Concepto de pipeline declarativo, ventajas vs Spark Structured Streaming manual*
	Sintaxis DLT	*@dlt.table, @dlt.view, LIVE.table_name, Python vs SQL*
	Calidad de Datos	*Expectations (@dlt.expect, warn, drop, fail), monitoreo de calidad*
	Tablas Bronze / Silver / Gold	*Medallion Architecture implementation con DLT*
	Streaming con DLT	*Auto Loader, readStream, Kafka integration, Trigger.AvailableNow*
	Orquestación con Workflows	*Jobs, Tasks, Dependencies, Retry policies, alertas por email/Slack*
	Lab 06	*Construir pipeline Medallion completo con DLT: Auto Loader → Bronze → Silver (con expectations) → Gold*

07	MLflow y Machine Learning en Databricks	2h 30m
	Databricks ML Runtime	*ML clusters, pre-installed libs (sklearn, XGBoost, TF, PyTorch)*
	MLflow Tracking	*mlflow.log_param, log_metric, log_artifact, autolog, UI del experimento*
	MLflow Model Registry	*Registro de modelos, staging, production, archivado, versioning*
	AutoML	*Generación automática de notebooks, métricas comparativas, best run*
	Model Serving	*Real-time inference endpoints, batch scoring con spark_udf*
	Lab 07	*Entrenar modelo clasificación, registrar en MLflow, promover a producción y llamar endpoint REST*

08	Producción, Seguridad y Optimización	3h 00m
	Unity Catalog	*Metastore, catálogos, schemas, permisos granulares, data lineage*
	Gobernanza de Datos	*Row-level security, column masking, Attribute-based access control*
	Secrets y Credenciales	*Databricks Secrets (CLI/API), Secret Scopes, integración con Key Vault/Secrets Manager*
	Monitoreo y Observabilidad	*Cluster logs, Query profiler, Ganglia metrics, Spark UI avanzado*
	Cost Management	*DBU tracking, Cluster policies, tagging de recursos, Photon ROI*
	CI/CD para Data Engineering	*Databricks Asset Bundles (DAB), dbx, GitHub Actions workflows*
	Integraciones Externas	*Azure Data Factory, AWS Glue, Event Hubs/Kafka, S3/ADLS/GCS*
	Lab 08	*Configurar Unity Catalog permissions, crear job con CI/CD via DAB, implementar row-level security*

	Ponente:	Triny Del Olmo
	Es una profesional de datos con más de 10 años de experiencia en Business Intelligence y Data Engineering. Ha trabajado en sectores como tecnología y telemática, desarrollando soluciones para áreas como finanzas, supply chain, ventas y telemetría. Cuenta con un perfil técnico-analítico sólido, con dominio de herramientas como Python, SQL, PySpark, Power BI y Tableau. Actualmente se desempeña como Data Engineer en Data Services y cuenta con la certificación Databricks Professional Data Engineer. Se distingue por su capacidad para construir soluciones de datos escalables y por traducir necesidades de negocio en implementaciones técnicas eficientes.	

Metodología y Recursos

Modalidad de Enseñanza

- 30% teoría conceptual con diagramas
- 70% práctica en Databricks Community Edition
- 8 laboratorios guiados (uno por módulo)
- Proyecto integrador final opcional

Materiales Incluidos

- Notebooks descargables por módulo
- Datasets de práctica (CSV, JSON, Parquet)
- Slides PDF por módulo
- Cheat sheets PySpark, Delta Lake, DLT

Inversión y promociones

Inversión

Promociones

Modalidad: Livestream

DURACIÓN: 20 HORAS
5 VIERNES: 16:00 -20:00 HRS.

INICIO: 29 DE MAYO

SESIONES POR ZOOM
Sesiones en vivo y además serán grabadas y se compartirán por TL con consulta 24/7

INVERSIÓN: 6,499 + IVA

EMPRESAS: 6X5 + 15% DE DESCUENTO
Aplica ***15% DE DESCUENTO EXALUMNOS RHCECAM o en la INSCRIPCIÓN DE DOS O MÁS PERSONAS.***

Pagos con TDC.
Aplica comisión del 5%

Cancelación alumno(a): Se podrá ceder su espacio a otra persona o considerar inversión a cuenta de otro curso de interés.
Cancelación RHCECAM: Reembolso de inversión al 100%

OTROS CURSOS PROGRAMAD

ACTUARÍA Y SEGUROS · CIENCIA DE DATOS Y BIG DATA · COMPUTACIÓN Y BASES DE DATOS · ESTADÍSTICA · FINANZAS · OTRAS ÁREAS

Facebook
@rhcecam

Instagram
@rhcecam

YouTube
Próximamente

RHCECAM CONSULTORÍA Y CAPACITACIÓN
rmes@rhcecam.com
sApp: +52 5548663933