Aprende a **procesar y analizar grandes volúmenes de datos** con PySpark en Azure Databricks, la plataforma líder para Big Data en la nube.

Este curso combina la configuración de Databricks en Azure con el procesamiento distribuido en PySpark, desde la ingesta y transformación de datos hasta la implementación de modelos de Machine Learning con MLlib y la automatización de pipelines bajo la arquitectura Medallón (Bronze, Silver y Gold).

Ideal para científicos, analistas, ingenieros de datos y desarrolladores que buscan dominar Big Data en un **entorno moderno, eficiente y colaborativo.**

SEDE

SAN PEDRO: Del Mall San Pedro, 300 mts Norte y 50 mts Oeste, Edificio Omala, 2do piso

(506) 4030-5024 / 8414 4646 growupcr.com info@growupcr.com /growupcr

TEMARIO:

Módulo 1. Fundamentos de Big Data y Databricks
• ¿Qué es Big Data y qué resuelve PySpark?
• Arquitectura de Apache Spark y entorno colaborativo de Databricks.
• Creación de cuenta en Azure, costos y configuración inicial.
• Notebooks en Databricks: primeros pasos en PySpark.

Módulo 2. Ingesta y Gestión de Datos en la Nube
• Conexión a Azure Data Lake Storage (ADLS).
• Montaje con claves, SAS y Secret Scope.
• Lectura de múltiples formatos (CSV, JSON, TXT, Parquet).

Módulo 3. Transformaciones y Consultas con PySpark
• DataFrames, SQLContext y Spark SQL.
• Operaciones con columnas: withColumn, select, drop.
• Filtrado, ordenamiento y conversiones de tipos.
• Funciones de ventana y consultas distribuidas.

Módulo 4. Agrupaciones y Estadística Descriptiva
• Uso de groupBy, agg, count y funciones de agregación.
• Cálculo de métricas descriptivas.
• Tablas dinámicas (pivot tables) y operaciones avanzadas.

Módulo 5. Optimización y Delta Lake
• Lectura/escritura Delta y Time Travel.
• Upserts, limpieza de datos y Z-Ordering.
• Estrategias de optimización para grandes volúmenes de datos.

Módulo 6. Preparación para BI
• Creación de tablas Gold para análisis empresarial.
• Casos de negocio: Employees, Customers, Products y Sales.
• Preparación de datasets para dashboards en Power BI/Tableau.

Módulo 7. Ingeniería de Características para Machine Learning
• Preparación de datos categóricos y numéricos.
• Codificación: StringIndexer, OneHotEncoder.
• Escalado y normalización (StandardScaler).
• Creación de pipelines de transformación en Spark.

Módulo 8. Machine Learning en Big Data con MLlib
• Introducción a MLlib y flujo de modelos en Spark.
• Modelos de regresión y clasificación.
• Evaluación de modelos: precisión, recall y AUC.

Módulo 9. Automatización y Orquestación de Pipelines
• Canalizaciones de datos (Bronze › Silver › Gold).
• Jobs y programación de ejecuciones en Databricks.
• Reducción de costos y buenas prácticas de operación.

INFORMACIÓN DEL CURSO:

Duración: 24 horas
Modalidad: Virtual – 100% en línea
Requisitos:
• Conocimientos básicos de Python y análisis de datos.
• Familiaridad con pandas y lógica de programación.
• No se requieren conocimientos previos de PySpark.
• SQL básico
• Cuenta Azure (se ofrece crédito inicial de $200)

Inversión:
Cupo mínimo: 5 personas
Inscripciones: 4030 5024 / 8414 4646

Al finalizar el curso con una nota mayor a 70 se entrega un certificado de aprovechamiento.

Observaciones:
* Sujeto a matrícula de un mínimo de personas.
* Este curso está respaldado por la Política de Calidad de Cursos Grow Up, más información en https://www.growupcr.com/politicadecalidad

CONSULTAS E INSCRIPCIONES:
(506) 4030 5024 / 8414 4646
info@growupcr.com
www.growupcr.com

growup
DATA ANALYTICS