WE ARE TECH | WAT

PROGRAMA DEL CURSO

CURSO APACHE SPARK / PYSPARK

X€ 20h Incluye Certificado WAT

Domina Apache Spark con Python para procesamiento distribuido de datos a escala. Aprende DataFrames, Spark SQL, streaming y optimización en Databricks con casos reales de producción.

PROGRAMA

UNIDAD 1: FUNDAMENTOS DE SPARK Y EL ENTORNO DE TRABAJO

Entenderás arquitectura de Spark: driver, workers y particiones. Configurarás PySpark local o Databricks Community Edition. Crearás DataFrames desde ficheros CSV, JSON y Parquet. Aplicarás transformaciones básicas: select, filter, withColumn y groupBy.

UNIDAD 2: SPARK SQL Y PROCESAMIENTO AVANZADO

Registrarás DataFrames como vistas temporales y consultarás con Spark SQL. Aplicarás JOINs y operaciones multi-tabla. Usarás funciones de ventana en PySpark para análisis avanzado. Procesarás datos de fechas, textos y estructuras anidadas.

UNIDAD 3: STREAMING, OPTIMIZACIN Y PRODUCCIN

Implementarás pipelines de Structured Streaming con PySpark. Aplicarás técnicas de optimización: particionamiento, caching y Adaptive Query Execution. Gestionarás datos con Delta Lake para transacciones ACID. Desplegarás jobs en Databricks o GCP Dataproc.

HERRAMIENTAS

- Apache Spark
- PySpark
- Databricks

QUÉ VAS A CONSEGUIR

- Entender arquitectura distribuida de Apache Spark
- Crear y manipular DataFrames con PySpark
- Implementar Spark SQL para análisis complejos
- Aplicar JOINs optimizados con broadcast hints
- Usar funciones de ventana para cálculos avanzados
- Procesar datos complejos: arrays, mapas y JSON
- Implementar streaming en tiempo real con Structured Streaming
- Optimizar queries con particionamiento y caching
- Usar Delta Lake para transacciones ACID en Spark

¿PARA QUIÉN ES ESTE CURSO?

Para data engineers, científicos de datos y especialistas que necesitan procesar big data distribuido a escala en la nube.

- Data engineers que crean pipelines batch y streaming
- Científicos de datos que procesan datasets grandes
- Especialistas en big data que usan Databricks
- Profesionales que migran de Pandas a Spark
- Consultores que implementan soluciones Spark
- Equipos que necesitan procesamiento distribuido en producción
- Especialistas en optimización de performance data

¿Preparado para dar el siguiente paso?

wearetech.es