Introducción

En la era del Big Data y la Inteligencia Artificial, las organizaciones necesitan herramientas que les permitan procesar grandes volúmenes de datos de manera eficiente, escalable y colaborativa. Databricks se ha consolidado como una de las plataformas más avanzadas para el procesamiento y análisis de datos, combinando el poder de Apache Spark con una infraestructura optimizada en la nube. Su enfoque unificado permite gestionar flujos de trabajo de ingeniería de datos, ciencia de datos y machine learning en un solo entorno, facilitando la colaboración entre equipos y reduciendo la complejidad operativa.

Databricks ofrece una experiencia de desarrollo intuitiva y flexible, integrando notebooks interactivos que permiten ejecutar código en Python, Scala, SQL y R de manera fluida y eficiente. Además, su compatibilidad con los principales proveedores de nube como AWS, Microsoft Azure y Google Cloud facilita la escalabilidad y optimización del procesamiento de datos en entornos empresariales.

Uno de los aspectos más destacados de Databricks es su capacidad para optimizar la ejecución de consultas y pipelines de datos, gracias a tecnologías como Delta Lake, que mejora la confiabilidad y el rendimiento en la gestión de datos estructurados y no estructurados. Esto permite a las empresas implementar soluciones de data lakes, data warehouses y data engineering sin las limitaciones de los sistemas tradicionales.

En este urso, exploraremos cómo desarrollar y optimizar flujos de trabajo con Databricks, abordando desde los fundamentos del procesamiento de datos con Apache Spark hasta la implementación de modelos de machine learning en producción. Los participantes aprenderán a crear y gestionar clusters, procesar grandes volúmenes de datos de manera distribuida y desplegar modelos de machine learning de manera eficiente, aprovechando las ventajas de una plataforma diseñada para la era del Big Data y la Inteligencia Artificial.

Objetivos del Taller

El objetivo de este Taller es proporcionar a los participantes una comprensión sólida y práctica sobre Databricks, permitiéndoles desarrollar, optimizar y desplegar flujos de trabajo de procesamiento de datos y machine learning en un entorno distribuido y escalable. A través de este Taller, los alumnos aprenderán a utilizar Apache Spark sobre Databricks para manejar grandes volúmenes de datos, ejecutar consultas eficientes y aplicar técnicas avanzadas de ingeniería de datos.

El Taller está diseñado para introducir a los participantes en las capacidades de Databricks como plataforma unificada para el análisis de datos, abarcando desde la configuración de clusters y la ejecución de notebooks hasta la implementación de modelos de machine learning. Se explorarán conceptos clave como el uso de Delta Lake para mejorar la confiabilidad de los datos, la integración con servicios en la nube y la automatización de pipelines de datos.

Con un enfoque práctico, los participantes trabajarán con notebooks interactivos en Databricks, aplicando técnicas de procesamiento de datos con Python y SQL. Además, se abordarán estrategias para optimizar la ejecución de consultas, administrar permisos y asegurar la gobernanza de los datos dentro de la plataforma.

Este Taller está dirigido a científicos de datos, ingenieros de datos y analistas que deseen aprender cómo aprovechar Databricks para desarrollar soluciones escalables de procesamiento y análisis de datos. Al finalizar el Taller, los participantes estarán preparados para diseñar pipelines de datos eficientes, integrar Databricks con entornos empresariales y desplegar modelos de machine learning de manera automatizada, maximizando el rendimiento y la escalabilidad en proyectos de Big Data e Inteligencia Artificial.

Contenidos

Introducción al Taller
- Presentación del Taller y objetivos de aprendizaje.
- Introducción al Big Data y su importancia en la industria.
- Apache Spark como motor de procesamiento distribuido.

Databricks
- ¿Qué es Databricks y cómo funciona?
- Diferencias entre Databricks y Apache Spark tradicional.
- Configuración y gestión de clusters en Databricks.

Introducción a los RDD en Spark
- Definición y estructura de los Resilient Distributed Datasets (RDDs).
- Creación de RDDs en Databricks.
- Transformaciones y acciones básicas en RDDs.

Transformaciones en un RDD en Spark
- Operaciones avanzadas en RDDs (map, filter, reduceByKey).
- Optimización del procesamiento distribuido con RDDs.

Acciones sobre un RDD en Spark
- Uso de acciones en RDDs (collect, count, first, reduce).
- Impacto de las acciones en la ejecución de Spark.

Aspectos Avanzados sobre RDDs
- Persistencia y particionamiento de RDDs.
- Optimización del rendimiento con Spark UI en Databricks.

Introducción a Spark SQL
- Creación y manipulación de DataFrames en Databricks.
- Ejecución de consultas SQL sobre grandes volúmenes de datos.

Spark SQL Avanzado
- Optimización de consultas con Catalyst Optimizer.
- Uso de UDFs (User Defined Functions) en Spark SQL.

Funciones en Spark SQL
- Funciones de agregación y transformación en Spark SQL.
- Aplicación de funciones de ventana y manipulación de datos.

Proyecto Final
- Desarrollo de un pipeline completo en Databricks.
- Aplicación de RDDs, Spark SQL y DataFrames.
- Evaluación y optimización del flujo de trabajo en Databricks.

Metodología

El Taller de Desarrollo con Databricks se desarrollará con una metodología enfocada en el aprendizaje práctico y progresivo, combinando teoría con aplicaciones en entornos reales de procesamiento de datos. Aproximadamente el 70% del Taller estará basado en ejercicios prácticos dentro de notebooks de Databricks, mientras que el 30% se enfocará en la teoría y los fundamentos de Spark y Databricks.

Cada módulo incluirá demostraciones en vivo, donde los participantes podrán ver en acción las capacidades de RDDs, Spark SQL y DataFrames, aplicándolos en escenarios reales dentro de Databricks. Los alumnos trabajarán en entornos en la nube y experimentarán con la gestión de clusters, ejecución de consultas distribuidas y optimización de procesamiento de datos.

El Taller incentivará la participación y la colaboración, permitiendo a los participantes resolver problemas en conjunto, discutir estrategias de optimización y aplicar buenas prácticas en Databricks. A lo largo del Taller, los alumnos realizarán desafíos prácticos y mini proyectos, que reforzarán los conceptos enseñados en cada módulo.

Al finalizar el Taller, los participantes trabajarán en un proyecto final, donde deberán construir un pipeline completo en Databricks, aplicando las técnicas de procesamiento de datos aprendidas. Este enfoque garantizará que los alumnos no solo comprendan los fundamentos, sino que también sean capaces de implementar soluciones escalables en un entorno real de Big Data.

Requisitos

Para aprovechar al máximo este Taller, es recomendable que los participantes cuenten con conocimientos previos en procesamiento de datos, programación y fundamentos de Big Data. Sin embargo, dado que es un Taller de desarrollo en Databricks, se explicarán los conceptos fundamentales antes de abordar las implementaciones más avanzadas.

Es deseable que los alumnos tengan experiencia básica en Python o Scala, ya que son los principales lenguajes utilizados en Databricks para la manipulación de datos y la implementación de algoritmos de machine learning. También será útil conocer SQL, ya que se trabajará con consultas estructuradas en Spark SQL y en la gestión de datos dentro de la plataforma.

Familiaridad con Apache Spark es un plus, pero no es obligatorio, ya que se introducirá su funcionamiento dentro de Databricks. Asimismo, es recomendable tener una comprensión básica de conceptos de almacenamiento y bases de datos, dado que trabajaremos con Delta Lake y la gestión de datos en entornos distribuidos.

Se recomienda contar con una cuenta en Azure Databricks, AWS Databricks o Google Cloud Databricks, ya que el Taller incluirá ejercicios prácticos en entornos de nube. También es conveniente tener conocimientos básicos sobre servicios en la nube y manejo de entornos de desarrollo en la web.

Este Taller está diseñado tanto para ingenieros de datos, científicos de datos y analistas que buscan optimizar su flujo de trabajo con Databricks, como para desarrolladores que deseen profundizar en arquitecturas escalables para Big Data e inteligencia artificial. La motivación por aprender y experimentar con herramientas de procesamiento distribuido será clave para aprovechar al máximo la formación.

Dirigido a:

Ingenieros de Datos, científicos de datos y analistas

Generalidades

- Duración 32 horas cronológicas.
- Taller cerrado.