**PREWORK**

Sesión 05

**Introducción a Spark**

Introducción a Spark

**🎯 Objetivo**

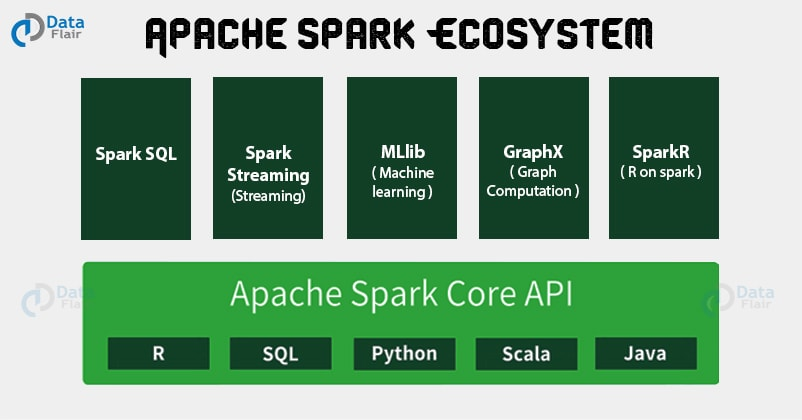
* Entender conceptualmente cómo se integra el Framework básico de Spark, y cómo se logra la administración y análisis de Big Data a través de la optimización de recursos y el uso indirecto del HDFS de Hadoop en el mismo Spark

### 

### 🚀 **Desarrollo**

### **Parte 1: ¿Qué es Spark?**

Apache Spark promete un procesamiento de datos más rápido y un desarrollo más sencillo de grandes cantidades de datos. ¿Cómo se logra Spark esto? Para responder a esta pregunta, debemos conocer a detalle el ecosistema de Apache Spark. Spark surge en buena parte debido a algunos problemas en particular que surgieron al usar Hadoop MapReduce.



I. Spark Core

Es el núcleo de Spark, que proporciona una plataforma de ejecución para todas las aplicaciones de Spark. Es una plataforma generalizada para admitir una amplia gama de aplicaciones.

ii. Spark SQL

Permite a los usuarios ejecutar consultas SQL / HQL en la parte superior de Spark. Con Apache Spark SQL, podemos procesar datos estructurados y semiestructurados. También proporciona un motor para que Hive ejecute consultas sin modificar hasta 100 veces más rápido en implementaciones existentes de Hadoop.



iii. Spark Streaming

Apache Spark Streaming permite una potente aplicación interactiva y de análisis de datos a través de la transmisión de datos en vivo. Las transmisiones en vivo se convierten en micro lotes que se ejecutan sobre el núcleo de Spark. Consulte nuestro tutorial de Spark Streaming para obtener un estudio detallado de Apache Spark Streaming.



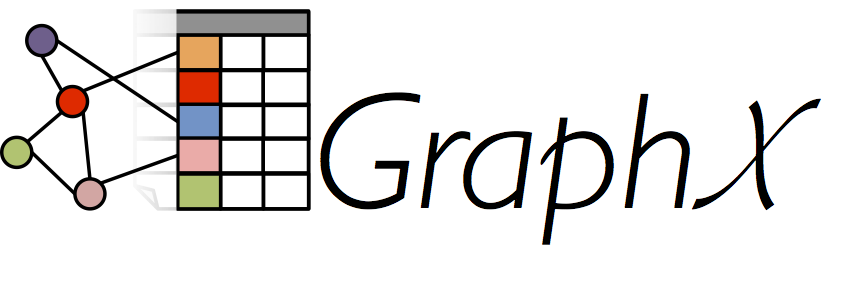
iv. Spark MLlib

Es la biblioteca de aprendizaje automático escalable que ofrece tanto la eficiencia como el algoritmo de alta calidad. Apache Spark MLlib es una de las mejores opciones para Data Science debido a su capacidad de procesamiento de datos en memoria, que mejora drásticamente el rendimiento del algoritmo iterativo.



v. Spark GraphX

Apache Spark GraphX ​​es el motor de cálculo de gráficos construido sobre Spark que permite procesar datos de gráficos a escala.



vi. SparkR

Es el paquete R que brinda una interfaz liviana para usar Apache Spark de R. Permite a los científicos de datos analizar grandes conjuntos de datos y ejecutar trabajos de manera interactiva en ellos desde el shell R. La idea principal detrás de SparkR era explorar diferentes técnicas para integrar la usabilidad de R con la escalabilidad de Spark.



vii. Resilient Distributed Dataset - RDD

Resilient Distributed Dataset (RDD) es la unidad fundamental de datos en Apache Spark, que es una colección distribuida de elementos en los nodos del clúster y puede realizar operaciones paralelas. Los RDD de Spark son inmutables, pero pueden generar nuevos RDD transformando los RDD existentes.

Hay tres formas de crear RDD en Spark:

* Colecciones paralelas: podemos crear colecciones paralelizadas invocando el método paralelizar en el programa controlador.
* Conjuntos de datos externos: al llamar a un método textFile, se pueden crear RDD. Este método toma la URL del archivo y la lee como una colección de líneas.
* RDD existentes: al aplicar la operación de transformación en RDD existentes, podemos crear nuevos RDD.

viii. Spark Shell

Apache Spark proporciona un Spark interactivo. Ayuda a que las aplicaciones Spark se ejecuten fácilmente en la línea de comandos del sistema. Usando el shell Spark podemos ejecutar / probar el código de nuestra aplicación de forma interactiva. Spark puede leer de muchos tipos de fuentes de datos para poder acceder y procesar una gran cantidad de datos.