Spark Stack (Pila de Spark)

El término "Spark Stack" (o Pila de Spark en español) se refiere generalmente a la **colección de componentes y bibliotecas principales que conforman el ecosistema de Apache Spark**. Es el conjunto integrado de herramientas diseñado para el procesamiento y análisis de Big Data.

Piensa en ello como una arquitectura en capas donde **Spark Core** forma la base, y varias bibliotecas especializadas se construyen sobre él para manejar diferentes tipos de tareas de procesamiento de datos.

Aquí tienes un desglose de los componentes clave que típicamente se incluyen en la Pila de Spark:

- **Spark Core:** Este es el **corazón** de Apache Spark. Proporciona las funcionalidades fundamentales para el envío, la planificación y las operaciones básicas de entrada/salida de tareas distribuidas. Es responsable de la gestión de la memoria, la tolerancia a fallos y la interacción con los sistemas de almacenamiento. Todas las demás bibliotecas de Spark se construyen sobre Spark Core. Se centra principalmente en el **procesamiento por lotes (batch processing)**.
- **Spark SQL:** Este componente está diseñado para trabajar con **datos estructurados y semiestructurados**. Proporciona una interfaz para consultar datos a través de SQL o una API de DataFrame (que es similar a las tablas en bases de datos relacionales o los data frames en R/Python). Spark SQL puede conectarse a diversas fuentes de datos (como Hive, bases de datos, archivos Parquet, etc.) y permite realizar consultas SQL y manipulaciones de datos en ellos de forma distribuida. También incluye **Spark Structured Streaming**, que es una extensión para manejar datos de transmisión en tiempo real (streaming) con operaciones similares a SQL y semántica de "exactamente una vez".
- **Spark Streaming:** Esta biblioteca permite el procesamiento de **flujos de datos en tiempo real (real-time data streams)**. Ingresa datos de diversas fuentes (como Kafka, Flume, Kinesis, sockets TCP) y los procesa en pequeños lotes (micro-batching). Spark Streaming permite aplicar las mismas transformaciones y operaciones de análisis a los datos de transmisión que se aplicarían a los datos por lotes. Un enfoque más nuevo y robusto para el procesamiento de flujos dentro de Spark es **Spark Structured Streaming**, que se construye sobre Spark SQL.
- **MLlib (Machine Learning Library):** Esta es la biblioteca de aprendizaje automático escalable de Spark. Proporciona una amplia gama de algoritmos distribuidos para tareas comunes de aprendizaje automático, tales como:
    - **Clasificación:** Predicción de etiquetas categóricas (por ejemplo, detección de spam).
    - **Regresión:** Predicción de valores continuos (por ejemplo, predicción del precio de viviendas).
    - **Clustering (Agrupamiento):** Agrupación de puntos de datos similares (por ejemplo, segmentación de clientes).
    - **Filtrado Colaborativo:** Construcción de sistemas de recomendación.
    - **Ingeniería de Características:** Herramientas para la extracción, transformación y reducción de dimensionalidad de características.
    - **Pipelines (Tuberías):** Utilidades para construir, evaluar y ajustar flujos de trabajo de aprendizaje automático.
- **GraphX:** Esta es la biblioteca de Spark para el **procesamiento de grafos**. Proporciona APIs para manipular grafos (redes de vértices y aristas) y realizar cálculos paralelos en grafos. Es útil para analizar relaciones y estructuras en los datos, como redes sociales, grafos de recomendación y análisis de redes.