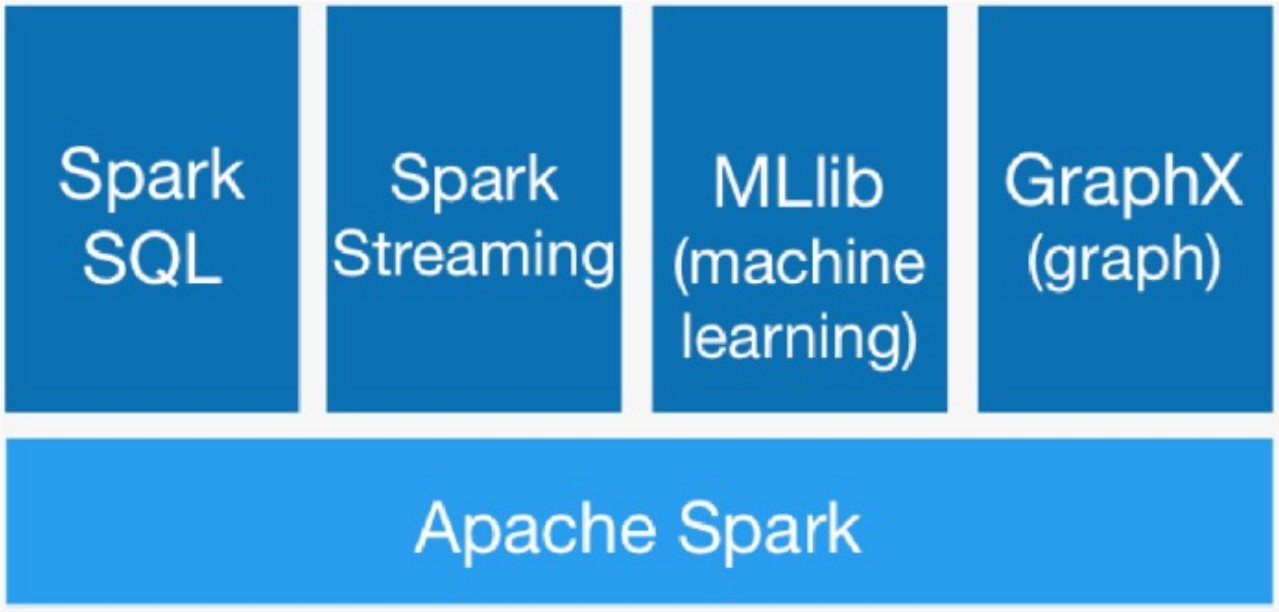
Módulos de Spark

Apache spark es un proyecto que está compuesto por diferentes herramientas y librerías,que se complementan unas con otras con el fin de ofrecer una solución completa para el procesamiento y análisis de big data.

Apache Spark se compone principalmente de 4 módulos:



- **Spark Core:** constituye el núcleo de spark,ofreciendo como principal abstracción el RDD(conjunto de datos distribuido resistente).Spark Core proporciona numerosas operaciones que se pueden realizar sobre los conjuntos de datos ,incluyendo operaciones mapReduce
- **Spark SQL o Dataframes,** que es un módulo que nos permite lanzar consultas en SQL o trabajar con el formato de dataframes incorporado del lenguaje de programación R.
- **Spark Streaming, módulo que se encarga de procesar datos en tiempo real.** Mientras MapReduce solo procesa datos en lotes, Spark tiene la posibilidad de gestionar grandes datos en tiempo real. Esto facilita que los datos se analicen según van entrando, sin tiempo de latencia y a través de un proceso de gestión en continuo movimiento.
- **Spark MLlib: Módulo especializado en funciones de machine learning.** Es una librería de spark que permite ejecutar tareas de aprendizaje automático. Esto permite aprender modelos a partir de datos que a continuación se pueden emplear para realizar clasificación de datos o recomendación.
- **Spark GraphX : módulo especializado en funciones de búsquedas en grafos**. Permite ejecutar operaciones sobre grafos(redes) de forma distribuida,permitiendo por ejemplo calcular el pageRank de una serie de sitios web a partir de la estructura de los mismos.