# Resumen 3

# Dillan Almendares Barrantes 2020033336

---

## Apache Spark

Fue diseñado para tener un motor de datos distribuidos unificado para solucionar un problema del big data, la necesidad de usar varios tipos de procesadores de datos, para lo cual tiene una extensión llamada “Resilient Distributed Datasets” (RDDS) para procesar varias tareas sin utilizar otros motores como: SQL, streaming, machine learning y graph processing. De manera que se alcanza un rendimiento similar por tener la optimización de un motor dedicado para cada tipo pero siendo ejecutadas como librerías sobre un mismo motor. Las ventajas que tiene el uso de un mismo motor son:

- Facilita el desarrollo de aplicaciones por utilizar una API unificada.
- Es más eficiente combinar tareas de procesamiento porque puede ejecutar funciones directamente en los datos, incluso en memoria sin tener que escribir datos en almacenamiento para pasarlos a otro motor.
- Permite el uso de ciertas aplicaciones como consultas interactivas en gráficos.

Desde su lanzamiento en 2010 Spark se ha convertido en el mayor proyecto open source para el procesamiento de big data, esto ha impulsado el desarrollo de una librería estandar integrada en Spark con funciones desde importar datos hasta de machine learning.

## Modelo de programación

La clave de la abstracción de Spark está en RDDs que son colecciones de objetos con tolerancia a fallas que están particionados en un cluster que pueden ser manipulados en paralelo. Son creados con las operaciones llamadas "transformaciones" (map, filter, groupBy) sobre los datos. Los RDDs son accesibles por una API donde los usuarios pueden pasar las funciones a ejecutar en el cluster. Las transformaciones retornan el RDD pero no son computadas en el momento sino que espera el llamado de una acción para ver todas las transformaciones y crear un plan de ejecución, donde si hay varias operaciones sobre una misma fila se combinan.

### Tolerancia a fallas

Los sistemas distribuidos suelen usar replicación de datos o checkpoints como método de protección ante fallas. En el caso de Spark se usa algo llamado "lineage" que consiste en que cada RDD recorra todas las transformaciones que fueron usadas en su elaboración, así retorna el resultado de cada operación en los datos para recuperarlos si se perdió alguna partición. Este sistema resulta más eficiente para datos grandes porque no hay que escribir a través de la red sino que en la RAM del nodo.

### Librerías de alto nivel

Con el modelo de RDD donde se tienen colecciones de objetos y funciones que ejecutarles, se diseñaron librerías para Spark que sirvan para aplicar técnicas usadas por otros motores para mejorar el rendimiento, las 4 librerías principales son:

1. SQL y DataFrames: Con Spark SQL se pueden implementar consultas relacionales, soportando almacenamiento columnar y generación de código para ejecutar consultas. Además agrega un nivel

más alto de abstracción en las transformaciones de datos llamado DataFrame que son RDDs con un esquema definido y con métodos definidos para filtrar, generar nuevas columnas y agregaciones.
2. Spark Streaming: Permite procesar streams incrementalmente, la implementación cortar el stream en pequeños lapsos (como de 200ms) que se combinan con lo almacenado en los RDDs para generar nuevos resultados.
3. GraphX: Provee de una interfaz para la computación de gráficos similar a Pregel y GraphLab, con las mismas optimizaciones.
4. MLlib: Es la librería de machine learning, tiene implementados más de 50 algoritmos para el entrenamiento distribuido de modelos.

Como todas las librerías trabajan con la abstracción de datos de los RDDs, es sencillo usar varias librerías en conjunto dentro de una aplicación, el rendimiento tampoco se ve perjudicado al utilizar las librerías sino que llega a ser comparable al de motores especializados.

# Aplicaciones

Spark es utilizado en más de mil compañías y en varios ámibitos como: servicios web, biotecnología y finanzas. Los principales usos que se le da son:

- Batch processing: El mayor uso que se le da a Spark es el batch processing de grandes datasets, también en convertir raw data a un formato estructurado y en el entrenamiento offline de modelos de machine learning.
- Consultas interactivas:
    1. Compañías usan Spark SQL para consultas relacionales, generalmente con herramientas de business intelligence como Tableau.
    2. Usando interfaces interactivas mediante shells para responder preguntas avanzadas y diseñar modelos para aplicaciones de producción.
    3. Varios proveedores han desarrollado aplicaciones de dominio específico que corren en Spark.
- Stream processing: Tanto para análisis y decisiones en tiempo real entre servidores.
- Aplicaciones científicas: Usado en detección a gran escala de spam, procesamiento de imágenes y datos genómicos.