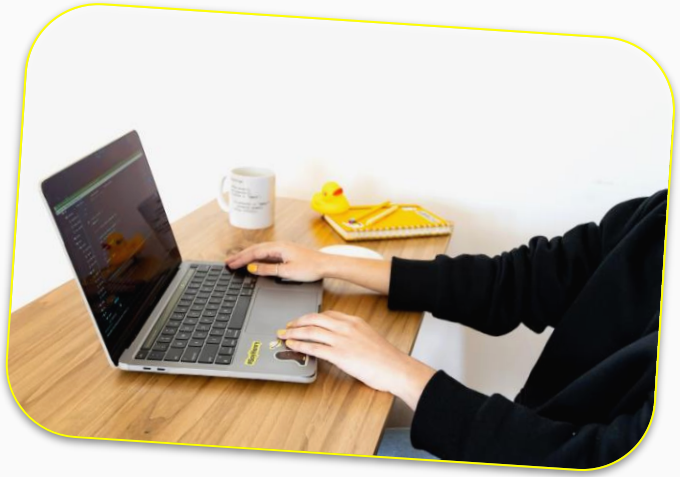
HENRY

Apache Spark

OBJETIVOS DE CLASE

- **Comprender** el caso de uso de Spark y su Arquitectura.
- **Entender** qué es el RDD -Conocer los Módulos de Spark.
- **Diferenciar** las características del Procesamiento Batch y del Procesamiento Streaming.
- **Aplicar** el caso de uso de Kafka.



AGENDA

- Apache Spark
  - Arquitectura Spark
  - Cluster Spark
  - Hadoop Spark
- RDD
- Módulos Spark
- Kafka

Apache Spark

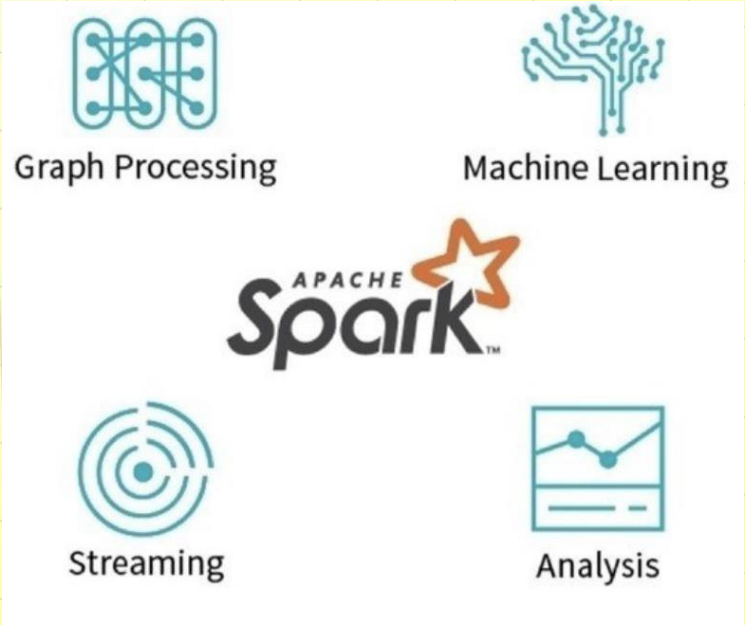
Introducción

¿Qué es?

Es un motor para procesamiento a gran escala de datos, integrado, rápido, "in memory" y de propósito general. Tienen su propio sistema de "Cluster Management" y utiliza Hadoop solo como almacenamiento. Spark está escrito en Scala y provee APIs en Java, Scala, Python y R.

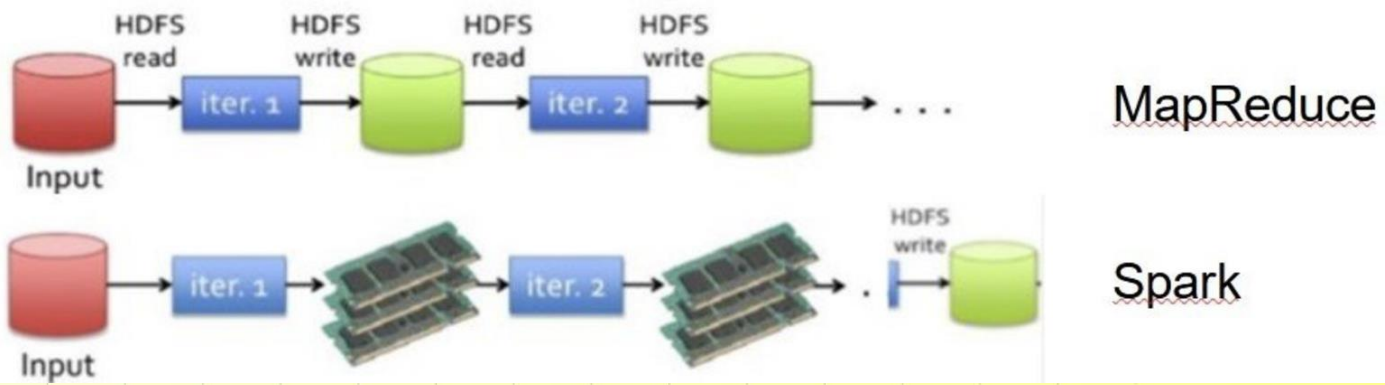
Características

- Computación en memoria.
- Tolerancia a fallos.
- Multipropósito.



¿Para qué sirve?

Spark es ideal para tareas de **procesamiento iterativo e interactivo de grandes "data sets" y flujos de datos ("streaming")**. Brinda una performance entre 10-100x mayor que Hadoop operando con construcciones de datos ("data constructs") llamadas "Resilient Distributed Datasets" (RDDs), esto ayuda a evitar latencias en tareas de lectura y escritura en discos. Es una alternativa a MapReduce.



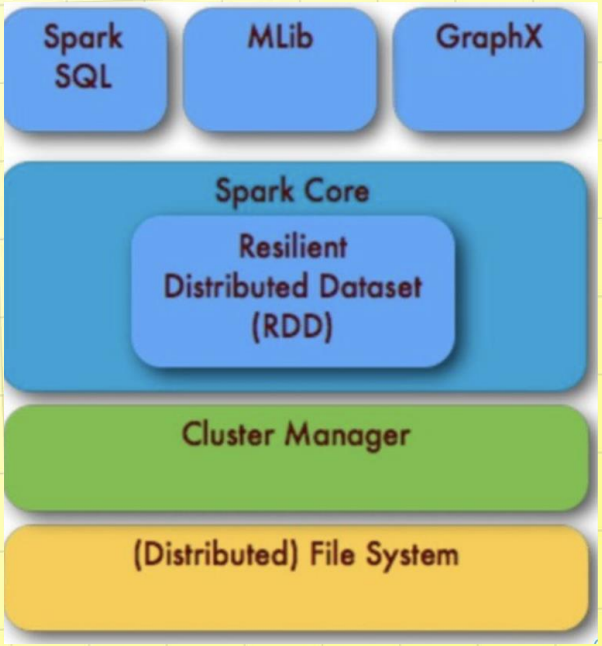
Arquitectura Spark

¿Para qué sirve?

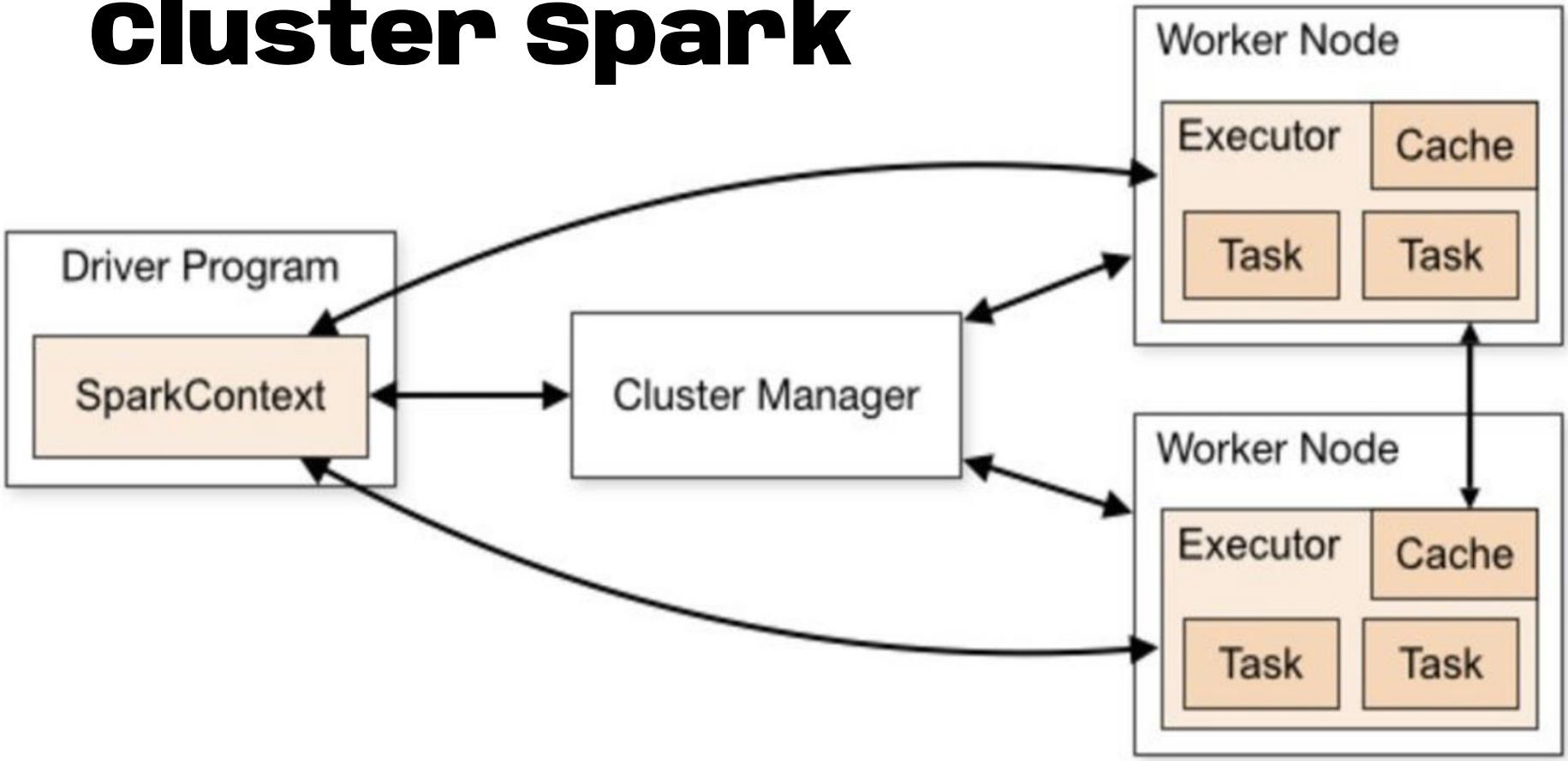
Spark tiene una arquitectura de capas bien definida donde todos los componentes están relacionados e integrados con extensiones y librerías.

Está basado en **dos abstracciones**:

- **RDD** (Resilient Distributed Dataset)
- **DAG** (Directed Acylic Graph)



Cluster Spark



Hadoop-Spark

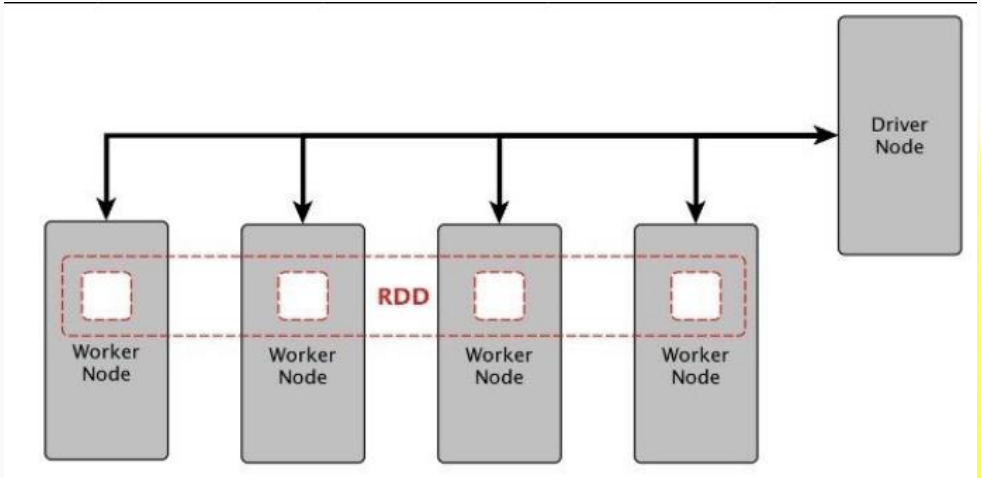
	Hadoop	Spark
Propósito	Procesamiento y almacenamiento de grandes datasets	Motor de propósito general para procesamiento de datos a gran escala.
Componentes principales	Hadoop Distributed File Systems	Spark core, motor de procesamiento en memoria.
Almacenamiento	HDFS administra colecciones de datos a través de múltiples nodos de un cluster de servidores “comodity”	Spark no realiza almacenamiento distribuido, opera en colecciones de datos distribuidas.
Tolerancia a fallos	Replicación. Los datos son escritos a discos en varios nodos luego de cada operación.	RDD’s minimizado “network I/O”. Los RDDs pueden ser reconstruidos ante fallos.
Velocidad de procesamiento	MapReduce es más lento.	Hasta 10x más rápido que MAPreduce para “batch processing” y hasta 100x más rápido para steaming processing.
Soporte de lenguajes	Java	Scala, APIs para Python, Java, R y otros.

RDD

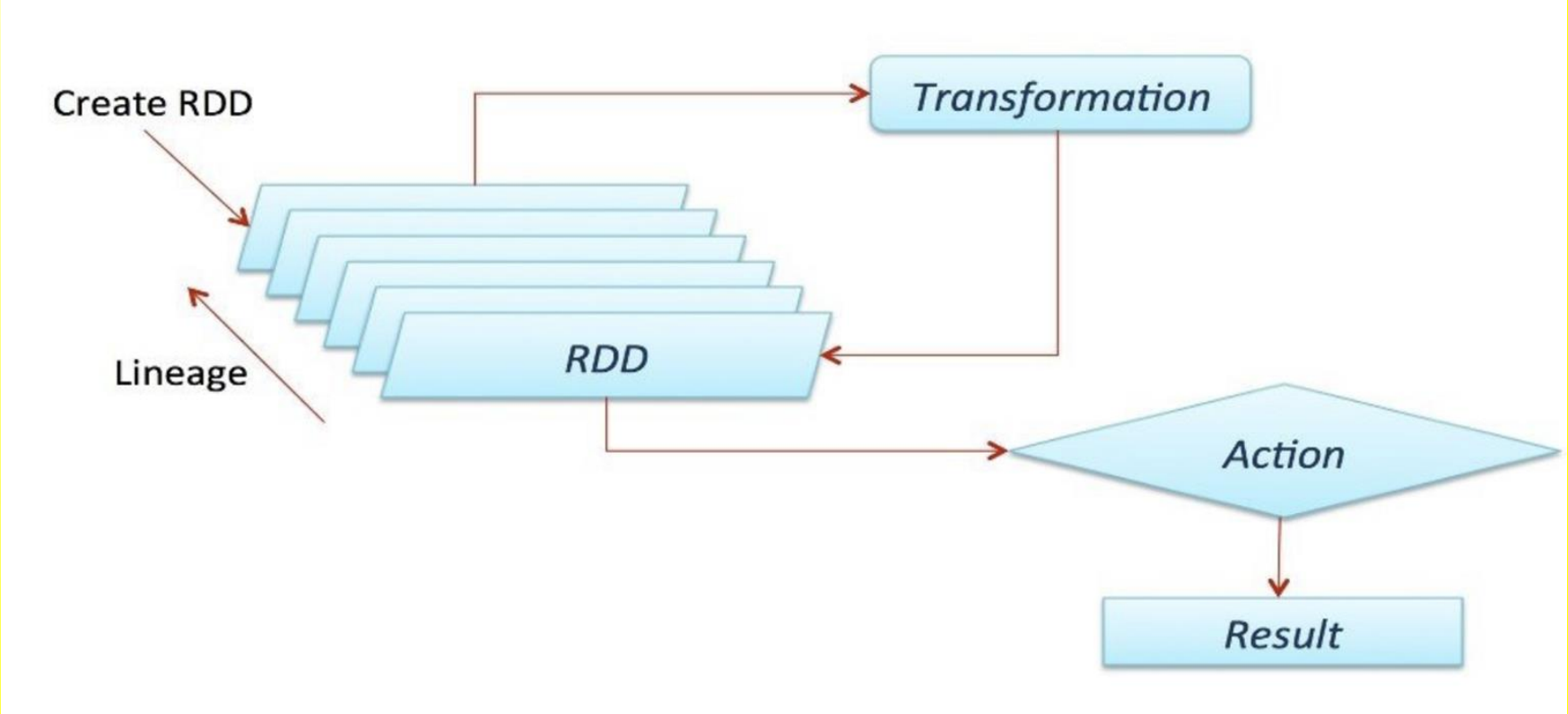
¿QUÉ ES?

- Es la estructura fundamental de datos de Apache Spark, una colección de objetos que se computan en diferentes nodos del Cluster.
- Resilient (tolerante a fallos), capacidad de recomponer particiones de datos dañadas o perdidas por fallos en nodos.
- Distributed, los datos residen en varios nodos.

- DataSet, representa registros de los datos, que el usuario puede cargar en forma de archivos JSON, CSV, texto o bases de datos por medio de JDBC sin una estructura de datos específica.



OPERACIONES SOBRE RDD

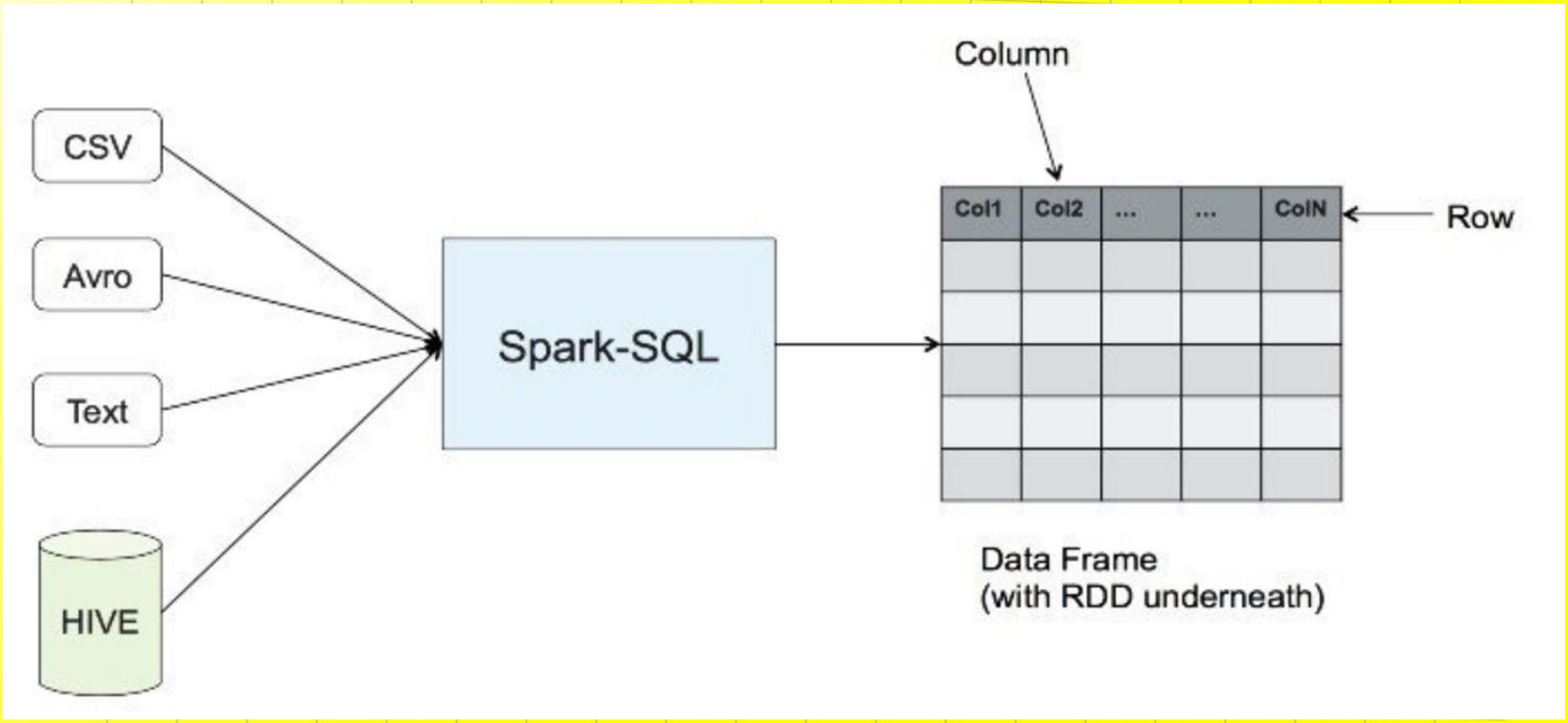


Dataframes

Dataframes

Colección de RDD’s con esquema. Características:

- Los datos están organizados en columnas nombradas.
- Es equivalente a una tabla en una base de datos relacional.

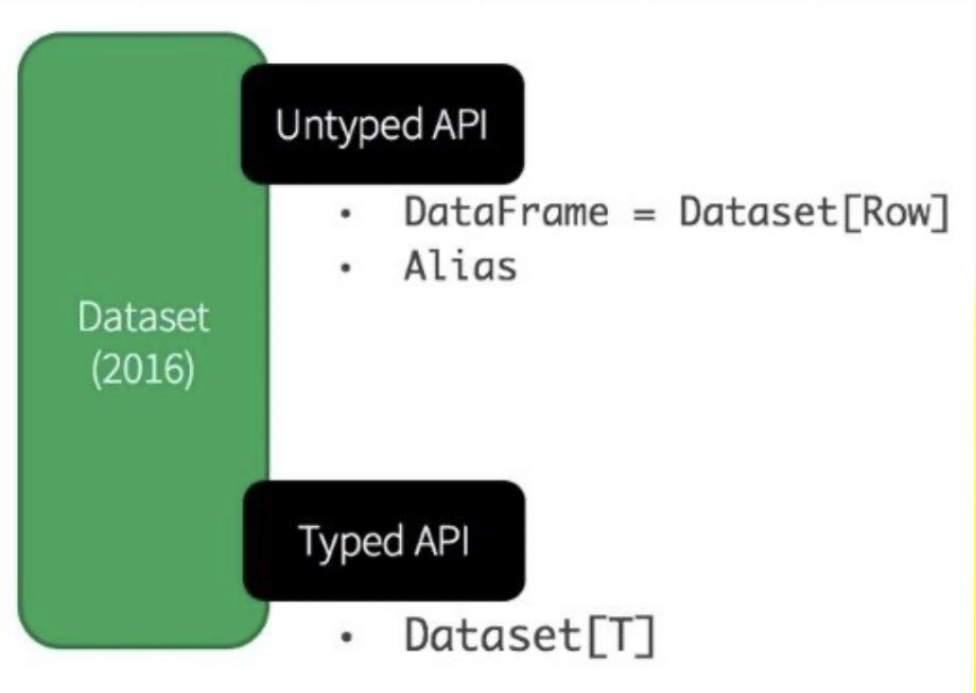

Column
CSV
Avro
Text
HIVE
Spark-SQL
Col1
Col2
...
...
ColN
Row
Data Frame
(with RDD underneath)

Datasets

¿QUÉ ES?

Es una extensión del Dataframe.
Características:

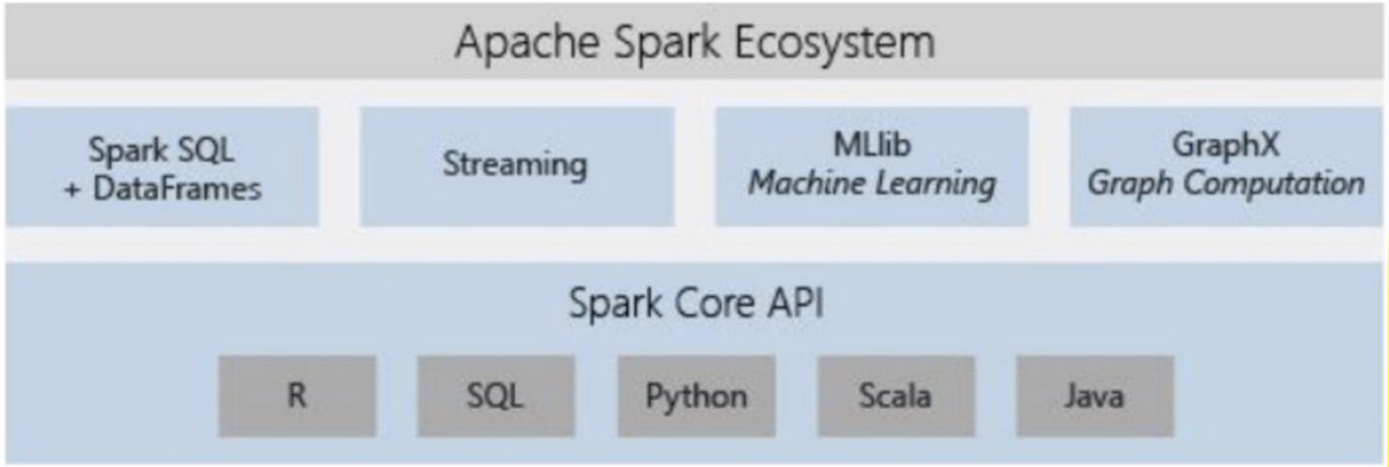
- Clases fuertemente tipadas.
- Verificación de tipos de dato en tiempo de compilación.
- Disponible sólo en Scala y Java.



Módulos Spark

Spark Core

- Brinda velocidad dando capacidades de computación ”in-memory”. Spark Core es la base del procesamiento distribuído de grandes datasets.



SparkSQL

- Lenguaje provisto para tratar con datos estructurados.
- Permite ejecutar consultas sobre los datos y obtener resultados útiles.
- Soporta consultas a través de SQL y HiveQL.

Spark Streaming

- Permite procesamiento de flujos en forma escalable, rápida y tolerante a fallos.
- Spark divide los “streams” de datos en pequeños “batches”.
- Trata a cada “batch” de datos como RDDs y los procesa.
- Puedo operar con varios algoritmos.

Spark MLlib

- Es una librería escalable de Machine learning.
- Contiene librerías para implementar algoritmos de ML, por ejemplo clustering, regression y Filtrado colaborativo.
- El workflow ML incluye estandarización, normalización, hashing, algebra lineal, estadísticas.

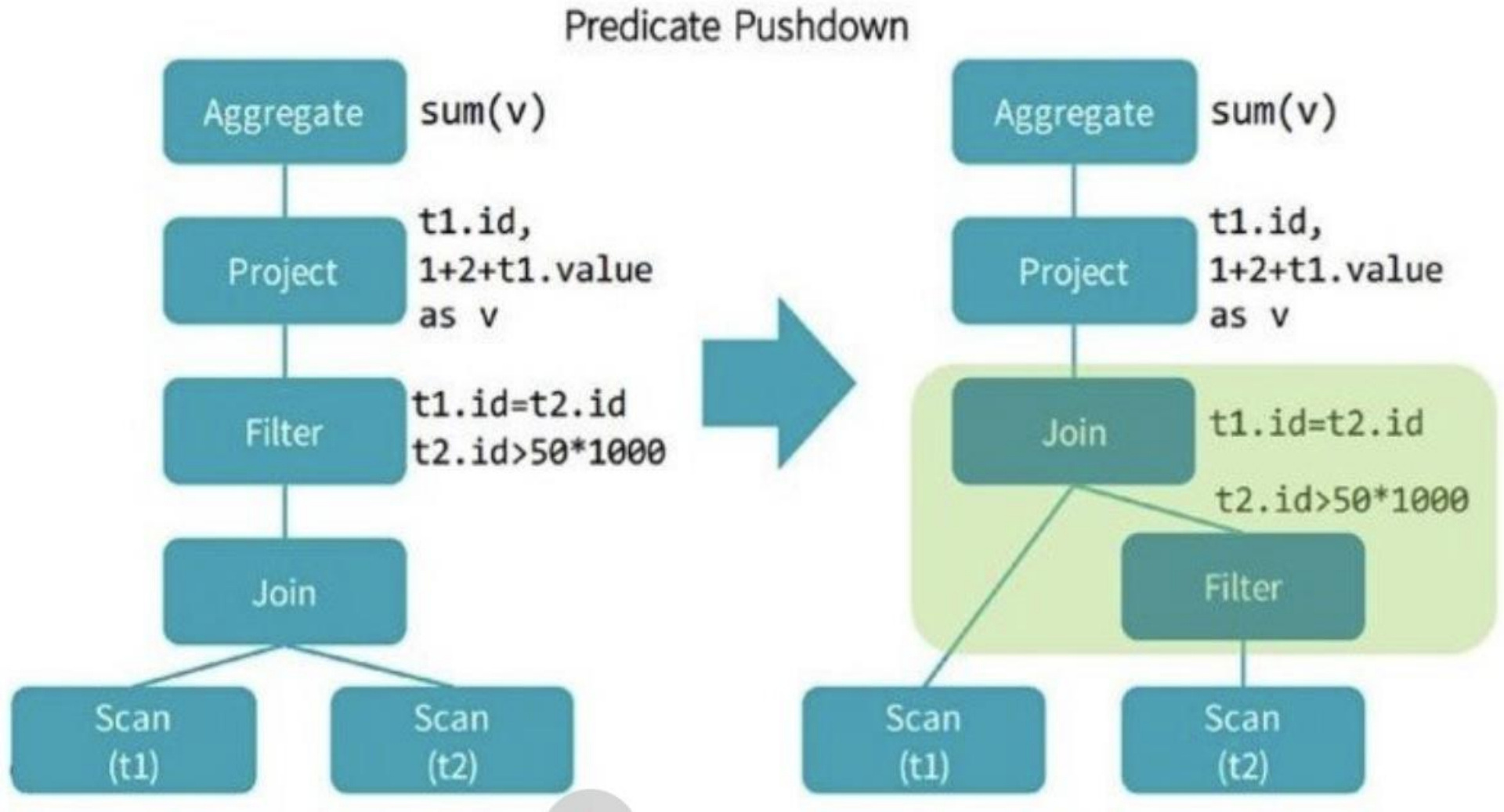
Spark GraphX

Es un motor de análisis de grafos. Extiende Spark RDD brindado una abstracción gráfica de grafos dirigidos con propiedades asignadas a cada nodo y vértice.

catalyst optimizer

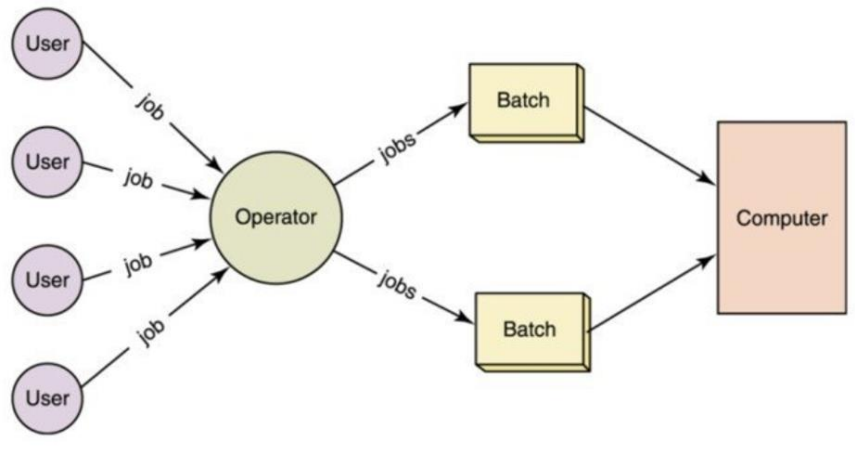
¿Qué es?

Motor de optimización de planes de ejecución, parecido al que usan las bases de datos, pero diseñado para la cantidad de datos propia de Spark. Además de eso, se tiene implementado un optimizador de memoria y consumo de CPU, llamado Tungsten, el cual determina cómo se convertirán los planes lógicos creados por Catalyst a un plan físico.


Predicate Pushdown
Aggregate
sum(v)
Project
t1.id,
1+2+t1.value
as v
Filter
t1.id=t2.id
t2.id>50*1000
Join
Scan
(t1)
Scan
(t2)
Aggregate
sum(v)
Project
t1.id,
1+2+t1.value
as v
Join
t1.id=t2.id
t2.id>50*1000
Filter
Scan
(t1)
Scan
(t2)

Procesamiento Batch

¿Qué es?



Es el procesamiento de transacciones por lote. Los trabajos que pueden ejecutarse sin la interacción del usuario final, o pueden programarse para ejecutarse según lo permitan los recursos.

Ej: Reporte anual de ventas.

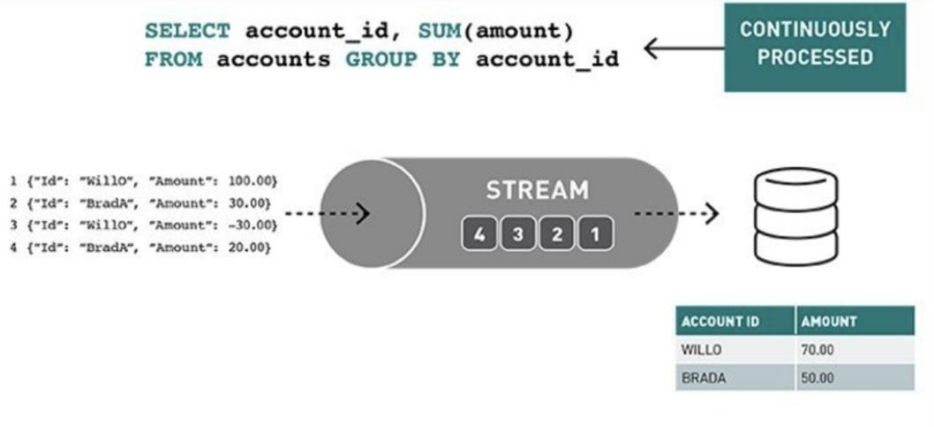
Procesamiento streaming

¿Qué es?

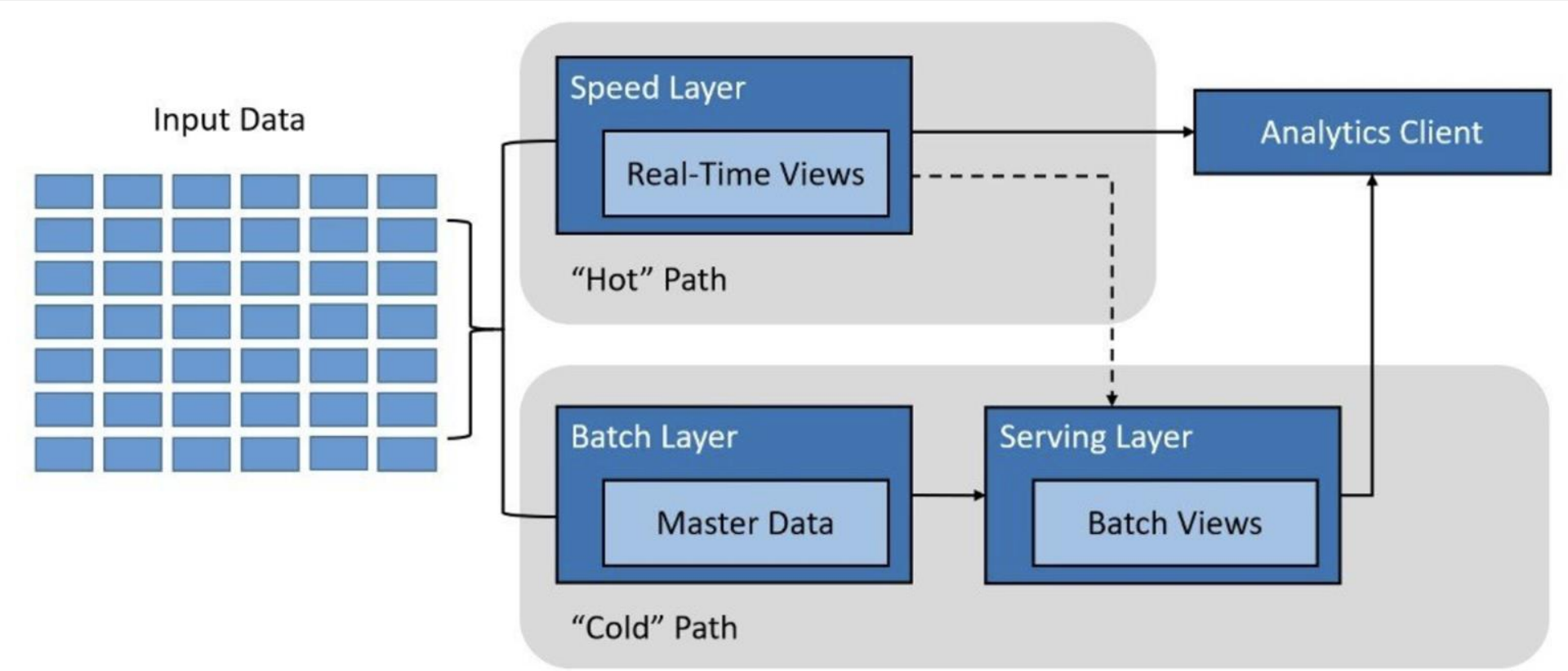
Procesamiento de datos a medida que se producen o reciben (flujo de datos en movimiento). Los datos se crean como una serie de eventos a lo largo del tiempo.

Ejemplos:

- Eventos de sensores
- Clicks en un sitio web
- Operaciones financieras



Arquitectura Lambda

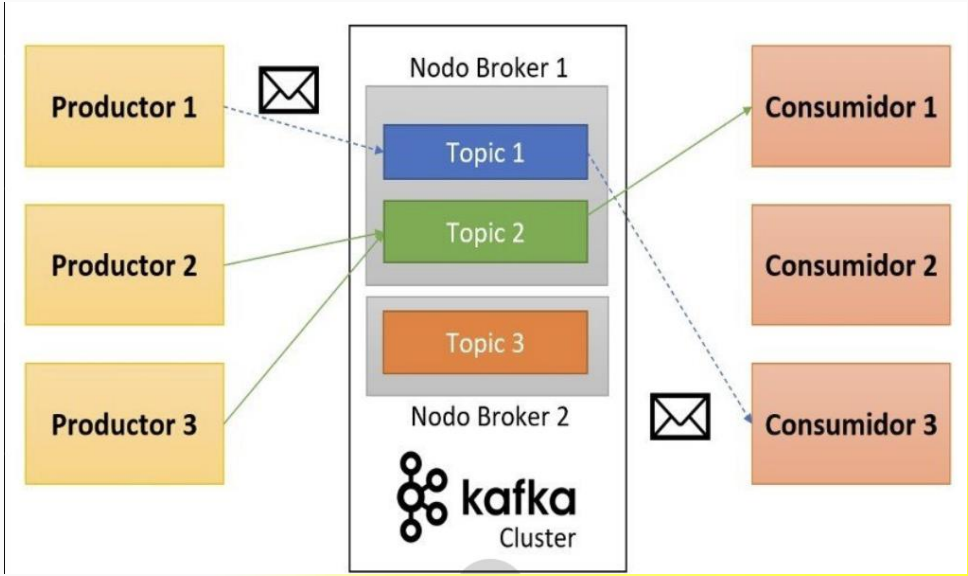

Input Data
Speed Layer
Real-Time Views
"Hot" Path
Analytics Client
Batch Layer
Master Data
Serving Layer
Batch Views
"Cold" Path

Kafka

¿Qué es?

Sistema de cola de mensajes distribuido que utiliza el patrón productor-consumidor

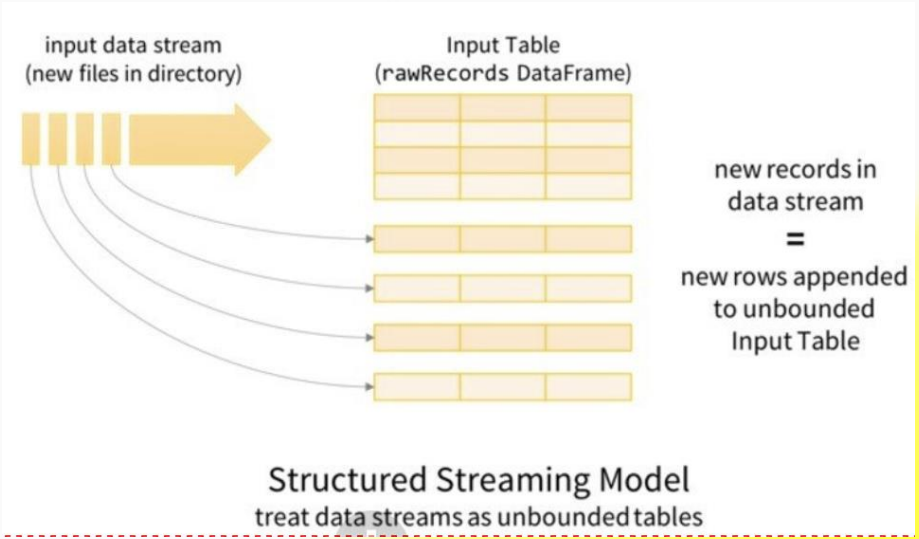
- Topics: categorías para los mensajes
- Producers: envían mensajes a un topic
- Consumers: se suscriben a tópics.
- Brokers: nodos que forman el cluster



spark streaming

¿Qué es?

- Es el módulo de Spark que nos permite ingestar y procesar flujos de datos continuos.
- Utiliza micro-batching para dividir flujos de datos continuos en batches correspondientes a un periodo de tiempo acotado.
- Posibles fuentes de datos : Kafka, Flume, AWS Kinesis, TCP Sockets, Twitter.



Dato

Apache Spark fue desarrollado originalmente en la Universidad de California, Berkeley, en el año 2009, como un proyecto de investigación denominado "AMPLab" (Laboratorio de Análisis de Datos en Memoria Distribuida). El objetivo principal era diseñar una plataforma de computación en clusters que pudiera procesar grandes volúmenes de datos de manera rápida y eficiente.

RESUMEN DE LA CLASE

- ✓ Spark
- ✓ Dataframes & Datasets
- ✓ Módulos sparks
- ✓ Kafka

¿PREGUNTAS?

¡Muchas gracias!