

# BIG DATA APLICADO



UT03: PROCESAMIENTO PARALELO  
EN APACHE HADOOP: SPARK

# ÍNDICE

---

- 1.- Fundamentos de Apache Spark**
- 2.- RDDs (Resilient Distributed Datasets)**
- 3.- Manipulación de datos con SparkSQL y DataFrames**
- 4.- Spark en IA y Machine Learning**



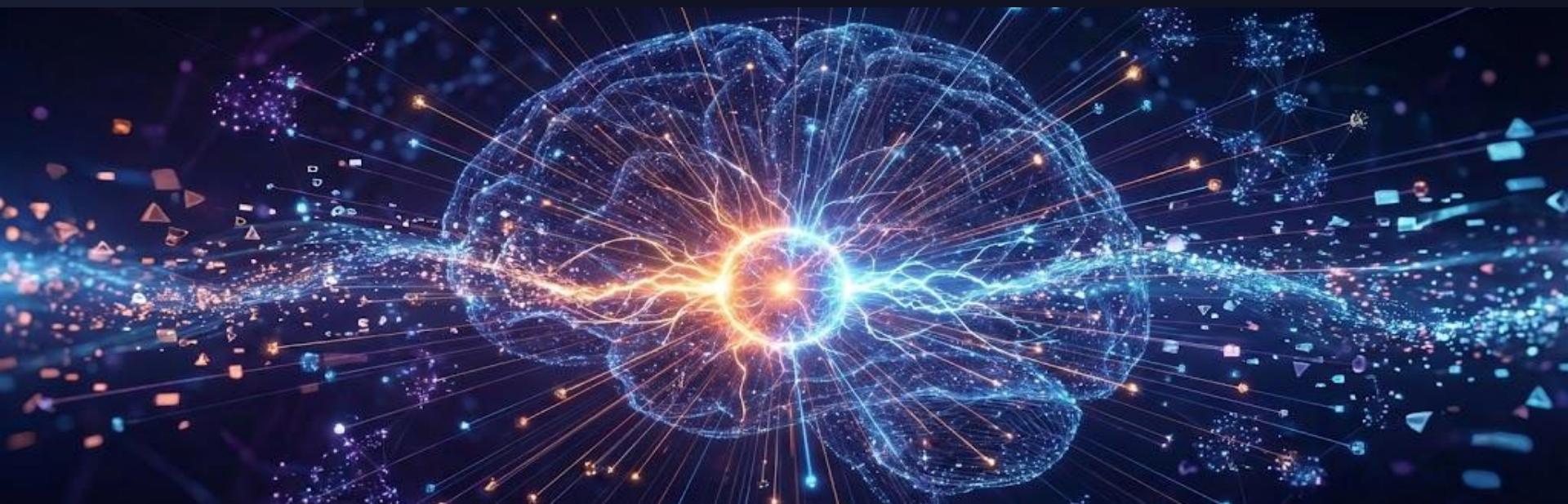
## Ejemplos Prácticos Recomendados

Para un curso práctico, es vital que implementen ejemplos de principio a fin.

- **Ejemplo 1:** ETL (Extract, Transform, Load) de grandes volúmenes de datos.
- **Ejemplo 2:** Construcción de un Sistema de Recomendación simple con MLlib.
- **Ejemplo 3:** Análisis de Sentimiento en tiempo real usando Structured Streaming y datos de ejemplo simulados (e.g., de Twitter/Kafka).

# 1

# FUNDAMENTOS DE APACHE SPARK



# 1

# FUNDAMENTOS DE APACHE SPARK

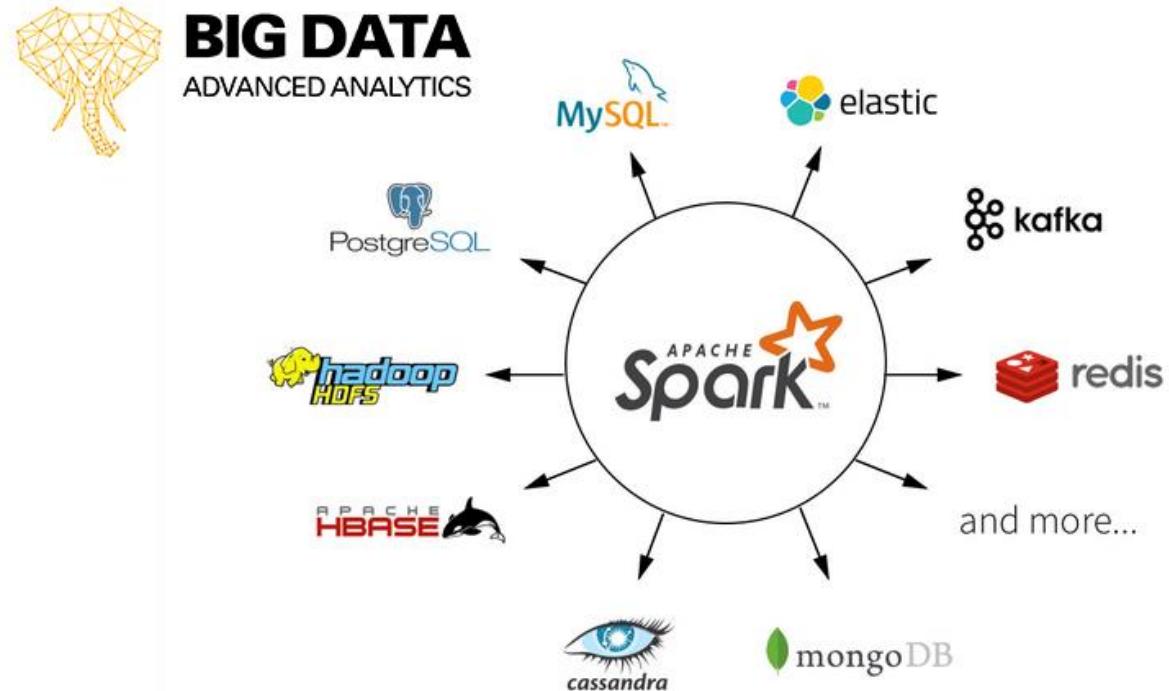
## 1.1

¿QUÉ ES APACHE  
SPARK?

Apache Spark es un **motor de análisis y procesamiento de datos distribuidos y de código abierto** diseñado para manejar grandes volúmenes de datos de manera eficiente y rápida.

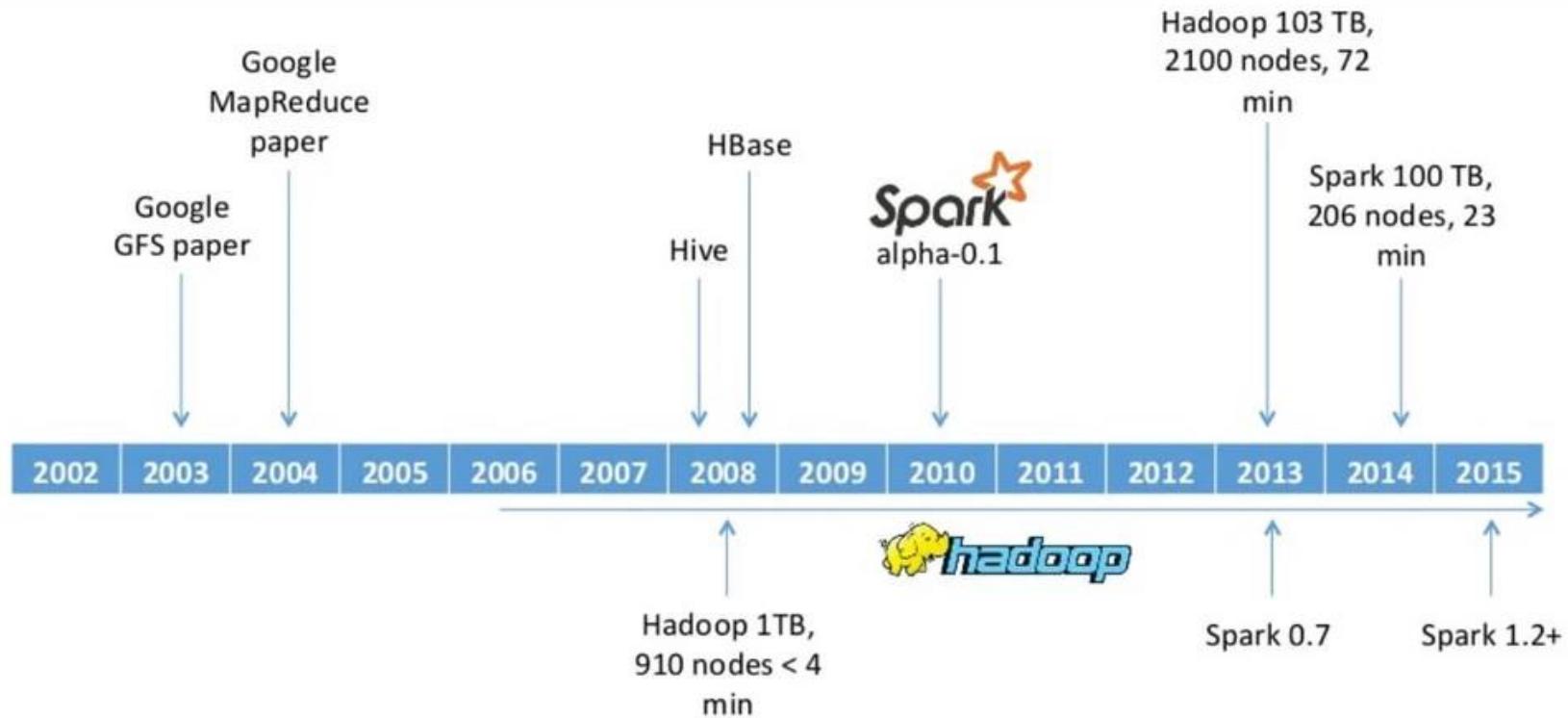
Su principal fortaleza es la capacidad de ejecutar tareas de **procesamiento en memoria**, lo que le hace extremadamente rápido.

Apache Spark puede leer datos de diversas fuentes (HDFS, S3, Cassandra, Kafka,...) y aplicar transformaciones complejas en paralelo a través de un clúster de máquinas.



## HISTORIA DE APACHE SPARK

- **2009:** nace como proyecto de investigación en el AMPLab de la Universidad de Berkeley, liderado por Matei Zaharia.
- **2010:** el código se publica como Open Source
- **2013:** se dona a la Apache Software Foundation
- **2014:** se convierte en un proyecto de nivel superior de Apache y rápidamente gana popularidad, superando a MapReduce como estándar de facto para el procesamiento de Big Data



## ARQUITECTURA DE APACHE SPARK

La arquitectura de Spark está diseñada en torno al concepto de procesamiento en un clúster de nodos.

### Driver Program (Programa Controlador)

Es el proceso que ejecuta el programa principal (p.e. un script de Python o Scala).

Contiene la **Spark Session** (o el antiguo **Spark Context**), que es el punto de entrada a toda la funcionalidad de Spark.

**Analiza, optimiza y distribuye la ejecución** del trabajo a los *Executors* en el clúster.

## Cluster Manager(Gestor del clúster)

Es un servicio externo (como **YARN**, **Mesos** o el propio **Spark Standalone**) que se encarga de adquirir recursos (cores y memoria) en el clúster.

Lanza los procesos *Executor* en los *Worker Nodes*.

## Worker Nodes (Nodos trabajadores)

Las máquinas físicas o virtuales **donde se ejecutan las tareas**.

Cada *Worker Node* aloja uno o más procesos *Executor*.

## Executors (Ejecutores)

**Los procesos que realmente ejecutan las tareas (tasks) de cómputo en los Worker Nodes.**

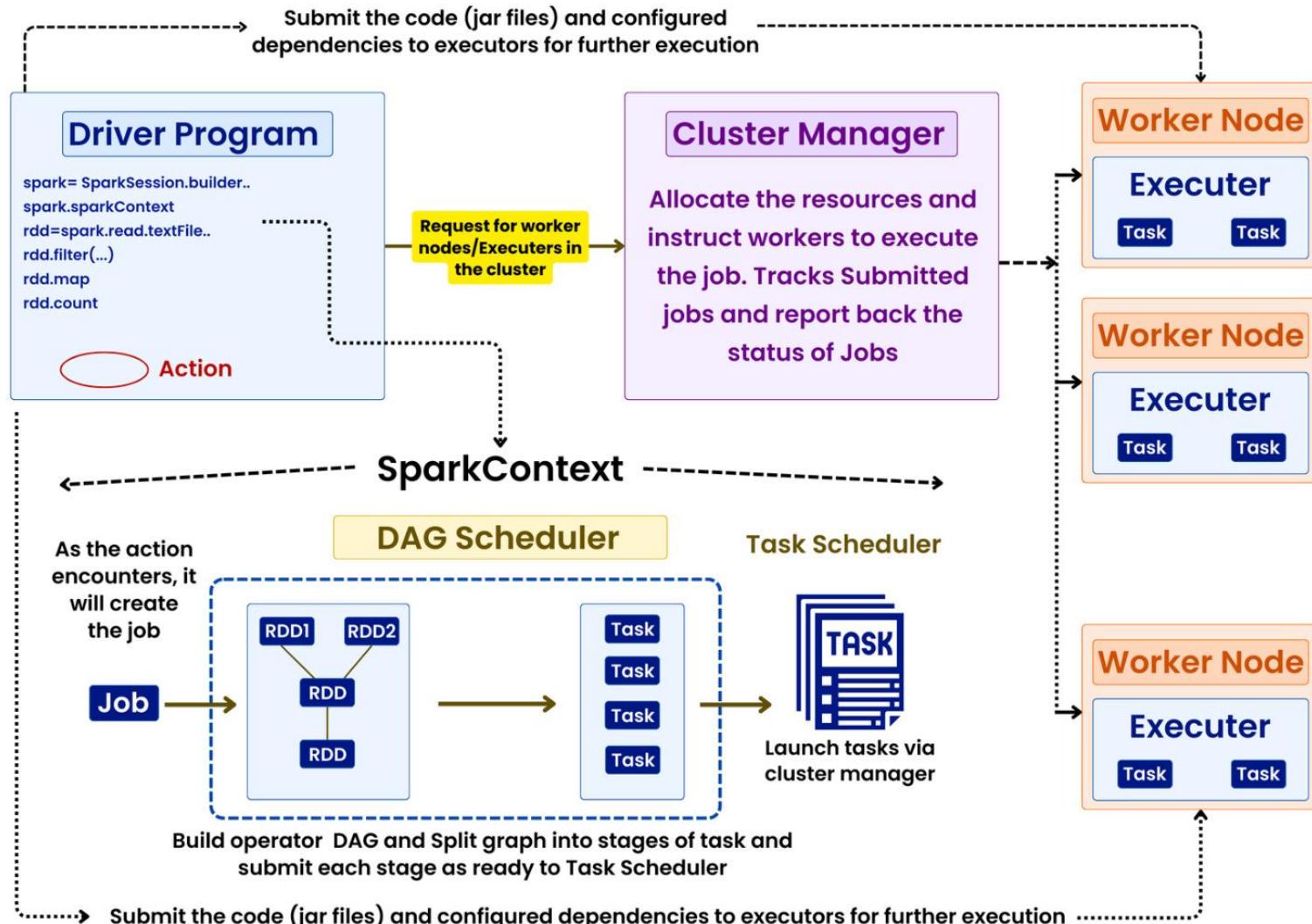
Almacenan los datos en caché de forma persistente.

Reportan el estado y los resultados de vuelta al *Driver Program*.

## Internals of Job Execution in Spark



Created by:  
Rocky Bhatia [in](#)



## RDD (Resilient Distributed Dataset)

Es una **colección inmutable, distribuida y tolerante a fallos de objetos** en el clúster de Spark, que se puede operar en paralelo.

Es la **abstracción** central de Spark Core

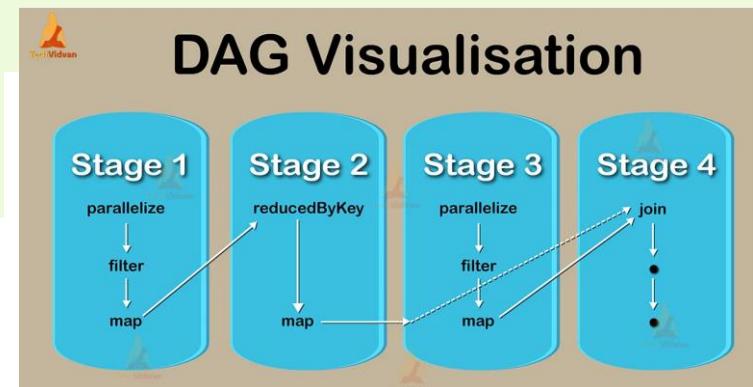
Sus principales características son:

### Resiliente (tolerante a fallos)

Un RDD es una **colección de elementos tolerante a fallos que se pueden operar en paralelo**.

Si un nodo falla, Spark puede reconstruir las particiones de datos perdidas desde el origen, sin perder el estado del cálculo.

Para ello utiliza los **Grafos Acíclicos Dirigidos (DAG)**



## Distribuido

Un RDD divide automáticamente los datos en varias piezas lógicas llamadas **particiones**.

Cada partición reside en un *worker* diferente del clúster y **se procesa en paralelo**.

El usuario escribe una operación (p.e. map) y Spark se encarga de dividir y orquestar miles de tareas en los diferentes nodos.

## Dataset (conjunto de datos)

El *dataset* describe el contenido del RDD

Los RDDs son colecciones de objetos (como una lista o array) que se encuentran dispersos en memoria o disco de las máquinas del clúster.

Los RDDs son **inmutables**. Una vez que se crea un RDD no se puede cambiar. Esto es esencial para la coherencia y la recuperación.

## Evaluación perezosa

Las operaciones con RDDs se dividen en dos categorías:

- **Transformaciones:**
  - Crean un nuevo RDD a partir del existente (map, reduce, ..)
  - No ejecutan nada inmediatamente, simplemente registran la operación en el DAG.
- **Acciones**
  - Son operaciones que activan el cómputo (collect, count, save, ...)
  - Cuando se llama a una acción, Spark mira el DAG y planifica la ejecución optimizada de todas las transformaciones

Las transformaciones definen la eficiencia del RDD.

Hay dos tipos:

- **Transformaciones estrechas:**

- Cada partición del RDD padre contribuye a una única partición en el RDD resultante (p.e. map, filter)
- Esto es rápido porque no requiere mover datos entre nodos

- **Transformaciones anchas:**

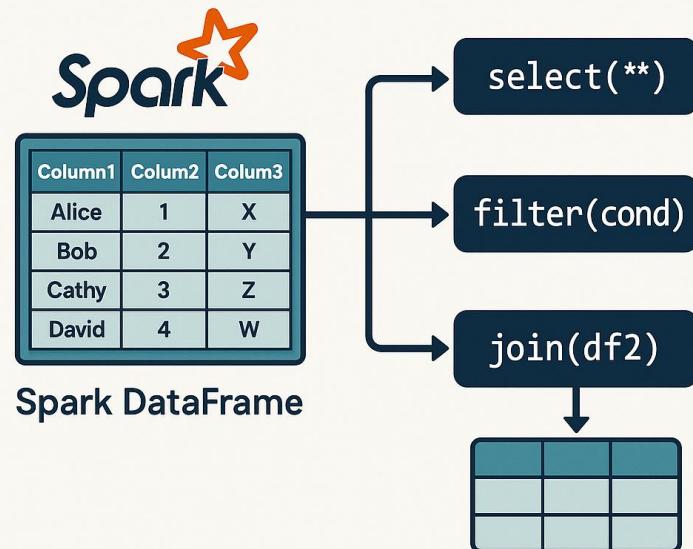
- Una partición del RDD padre puede contribuir a múltiples particiones en el RDD resultante (p.e. groupByKey, reduceByKey)
- Esto requiere un **shuffling**, que es la transferencia costosa de datos a través de la red del clúster

## Dataframes

Un dataframe es una **colección distribuida de datos organizados en columnas con nombre**

Conceptualmente es equivalente a una tabla de una base de datos relacional o a un *dataframe* de Pandas, pero con **optimizaciones masivas para ejecutarse en un clúster** de computadoras.

### Introduction to Spark DataFrames Basic DataFrame Operations



Sus principales características son:

- **Estructurados:** los dataframes tienen un esquema definido (frente a los RDDs que son colecciones de objetos arbitrarios)
- **Distribuidos:** los datos no residen en una sola máquina.
- **Inmutables:** una vez creado, un dataframe no se puede modificar, si se quiere cambiar algo hay que crear un nuevo dataframe a través de una transformación.
- **Evaluación perezosa:** al igual que pasa con los RDDs, Spark no ejecuta nada inmediatamente, sino que construye un plan lógico y solo procesa los datos cuando se llama a una acción (`show()`, `count()`, `write()`)
- **API unificada:** pueden ser manipulados desde diferentes lenguajes (Scala, Java, Python, R). También permiten ejecutar consultas SQL directamente sobre ellos creando vistas temporales.

# 2

# DATAFRAMES: INGESTA Y TRANSFORMACIÓN



**2.1**

**INTRODUCCIÓN A  
PYSPARK**

**2**

**DATAFRAMES:  
INGESTA Y  
TRANSFORMACIÓN**

Vamos a lanzar Spark en un entorno contenerizado

([https://vgonzalez165.github.io/docker\\_resources/compose/spark/](https://vgonzalez165.github.io/docker_resources/compose/spark/)) compuesto por un master, un worker y un tercer contenedor con Jupyter que nos servirá para programar con PySpark

●	<u>spark</u>	-	-	-
●	jupyter-spark-driver	e845bbfd37a2	<a href="#">spark-jupyter-driver</a>	<a href="#">8899:8888 ↗</a>
●	worker-1	62d4a6bb9690	<a href="#">spark:3.5.7-scala2.12-java11-pythc</a>	<a href="#">8081:8081 ↗</a>
●	master	0ee6e345dc6	<a href="#">spark:3.5.7-scala2.12-java11-pythc</a>	<a href="#">7077:7077 ↗</a> <a href="#">8080:8080 ↗</a>

## Conexión a Spark

Lo primero que tenemos que hacer es conectarnos a Spark

```
from pyspark.sql import SparkSession  
  
try:  
    spark = ( SparkSession.builder  
              .appName("Haciendo pruebas")  
              .master("spark://spark-master:7077")  
              .getOrCreate()  
    )  
  
    print("SparkSession iniciada correctamente.")  
except Exception as e:  
    print("Error en la conexión")  
    print(e)  
  
sc = spark.sparkContext
```

Comenzamos creando una sesión de Spark

Conectamos al contenedor master de Spark a través del puerto 7077

## Carga de datos en Spark

En este primer programa simularemos datos de prueba

```
data = [  
    ("Ana", 15),  
    ("Carlos", 22),  
    ("Luis", 10),  
    ("Marta", 35)  
]  
  
columns = ["nombre", "edad"]
```

Los dataframes tienen **estructura tabular**. En este caso crearemos una tabla de 2 columnas y 4 registros

Las columnas de las tablas tienen que tener un nombre

```
df = spark.createDataFrame(data, columns)
```

Creamos el dataframe

```
df.show()
```

```
+----+----+  
|nombre|edad|  
+----+----+  
| Ana | 15 |  
| Carlos | 22 |  
| Luis | 10 |  
| Marta | 35 |  
+----+----+
```

Con la función show() podremos mostrar el contenido de un dataframe

## Aplicando transformaciones

Vamos a aplicar nuestra primera **transformación** a los datos. Es importante entender que aquí Spark **no hace nada**, solo toma nota de lo que hay que hacer.

```
df_mayores = df.filter("edad >= 18")
df_mayores.explain()
== Physical Plan ==
*(1) Filter (isnotnull(edad#1L) AND (edad#1L >= 18))
+- *(1) Scan ExistingRDD[nombre#0,edad#1L]
```

Usamos **filter()** para filtrar solo los registros cuya edad sea mayor de 18 años

Con **explain()** podemos saber cuál es el plan de trabajo

Internamente Spark trabaja con **RDDs**.

Aquí vemos los pasos que Spark realizará para realizar la operación solicitada. A medida que vayamos añadiendo transformaciones realizará **optimizaciones** sobre este plan

## Acciones: ejecutando el plan de trabajo

Cuando realizamos una **acción** es cuando pone en marcha el plan de trabajo. En este caso vamos a usar la acción **show()**, que muestra los resultados

```
df_mayores.show()
```

```
+----+----+
|nombre|edad|
+----+----+
|Carlos| 22|
|Marta| 35|
+----+----+
```

# 2

# DATAFRAMES: INGESTA Y TRANSFORMACIÓN

## 2.2

### INGESTA DE DATOS

En el ejemplo anterior creamos los datos directamente en el código, pero lo habitual es obtenerlos de alguna fuente externa.

Spark puede recoger los datos desde diferentes fuentes como CSV, JSON o Parquet.

## Esquema

Algo importante antes de leer los datos es el **esquema**, que básicamente se puede definir como **el contrato que define la estructura de los datos**.

El esquema le dice tres cosas al motor sobre cada columna:

- **Nombre** (cómo se llama)
- **Tipo de datos** (qué es: DateType, IntegerType, ArrayType, ...)
- **Nullability** (¿puede estar vacío?)

Si no indicamos el esquema, Spark lo inferirá a partir de los datos, pero indicarlo explícitamente tiene una serie de ventajas:

- **Rendimiento:** si no lo definimos, Spark tendrá que leer el archivo completo antes de cargar los datos para intentar averiguar qué tipo de datos son
- **Seguridad de tipos:** evitar que una columna se cargue con un tipo erróneo (p.e. si en una columna de precios hay un registro con valor 10.5€ en lugar de 10.5)
- **Optimización:** Spark reserva memoria de forma más eficiente si sabe, por ejemplo, que una columna es Byte (1 byte) y no un Long (8 bytes)

Para definir el esquema, usaremos los siguientes objetos del módulo `pyspark.sql.types`:

- **StructType**: representa la **fila** completa. Es una colección (lista) de campos
- **StructField**: representa una **columna** individual

Los tipos de datos que hay son:

Equivalente en Python	Tipo PySpark	Descripción
<code>str</code>	<code>StringType()</code>	Texto libre
<code>int</code>	<code>IntegerType()</code>	Entero estándar (32 bits)
<code>int</code> (grande)	<code>LongType()</code>	Entero grande (64 bits)
<code>float</code>	<code>DoubleType()</code>	Decimales (por defecto en Spark)
<code>decimal</code>	<code>DecimalType()</code>	Finanzas (precisión exacta)
<code>datetime</code>	<code>TimestampType()</code>	Fecha y hora
<code>list</code>	<code>ArrayType(T)</code>	Lista de elementos de tipo T
<code>dict</code>	<code>MapType(K, V)</code>	Clave-valor

Veamos cómo sería el código para definir el esquema:

```
from pyspark.sql.types import StructType, StructField  
from pyspark.sql.types import StringType, IntegerType, BooleanType  
  
schema = StructType([  
    StructField("id", IntegerType(), False),  
    StructField("nombre", StringType(), True),  
    StructField("es_cliente", BooleanType(), True)  
])  
  
schema.fields  
  
[StructField('id', IntegerType(), False),  
 StructField('nombre', StringType(), True),  
 StructField('es_cliente', BooleanType(), True)]
```

Tenemos que importar los tipos de datos que usaremos

Instanciamos el objeto StructType.

Para cada campo definimos el nombre, el tipo de datos que contendrá y si puede ser nulo (True) o no (False)

Podemos usar la propiedad Fields para ver la estructura del esquema

Aunque el ejemplo anterior era una tabla plana, el esquema puede ser **anidado** conteniendo datos complejos (tipo JSON o NoSQL como MongoDB)

```
direccion_schema = StructType([
    StructField("calle", StringType(), True),
    StructField("cp", StringType(), True)
])
```

Creamos primero el esquema del tipo complejo

```
main_schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("nombre", StringType(), True),
    StructField("direccion", direccion_schema, True)
])
```

```
root
|-- id: integer (nullable = true)
|-- nombre: string (nullable = true)
|-- direccion: struct (nullable = true)
|   |-- calle: string (nullable = true)
|   |-- cp: string (nullable = true)
```

Y luego lo indicamos como tipo de datos del tipo principal

Para cargar datos directamente en el *dataframe* debemos utilizar la función **createDataFrame()**, a la que opcionalmente se le puede pasar el esquema.

```
data = [
    (1, "Juan Perez", ("Calle Gran Vía 25", "28013")),
    (2, "Maria Lopez", ("Av. Diagonal 100", "08018")),
    (3, "Carlos Ruiz", None)
]

df = spark.createDataFrame(data, schema=main_schema)
```

Usamos el parámetro `schema` para indicar el esquema al crear el *dataframe*

```
df.printSchema()
root
|-- id: integer (nullable = true)
|-- nombre: string (nullable = true)
|-- direccion: struct (nullable = true)
|   |-- calle: string (nullable = true)
|   |-- cp: string (nullable = true)
```

Con la función `printSchema()` podremos visualizar el esquema del *dataframe*

Si tengo campos anidados, podré acceder a ellos utilizando la notación del punto

La transformación  
select() filtra por  
columnas

```
df.select("nombre", "direccion.calle").show()
```

nombre	calle
Juan Perez	Calle Gran Vía 25
Maria Lopez	Av. Diagonal 100
Carlos Ruiz	NULL

La acción show()  
muestra el dataframe

Para cargar datos desde un fichero debemos utilizar el objeto **spark.read**, que nos da acceso a la API **DataFrameReader**.

La sintaxis general para leer un fichero es:

```
df = ( spark.read  
    .format("tipo_archivo")  
    .option("clave", "valor")  
    .schema(mi_esquema)  
    .load("ruta/al/archivo")  
)
```

Formato de la fuente de datos: csv, json, parquet, orc, text, jdbc

Configuraciones específicas del formato (p.e. separador para CSV o multiline para JSON)

Si hemos creado manualmente el esquema lo indicamos aquí

Ruta al fichero, puede ser local (`file://`) o distribuida (`hdfs://`, `s3a://`)

## Carga de datos de CSV.

Vamos a leer el fichero worldcities.csv

Name	Modified
recursos	6 min. ago
PR01_No_evaluable.ipynb	24 days ago
spam.data	24 days ago
Untitled.ipynb	24 days ago
Untitled1.ipynb	24 days ago
worldcities.csv	8 min. ago

	city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
1	Tokyo	Tokyo	35.6870	139.7495	Japan	JP	JPN	Tōkyō	primary	37785000	1392685764
2	Jakarta	Jakarta	-6.1750	106.8275	Indonesia	ID	IDN	Jakarta	primary	33756000	1360771077
3	Delhi	Delhi	28.6100	77.2300	India	IN	IND	Delhi	admin	32260000	1356872604
4	Guangzhou	Guangzhou	23.1300	113.2600	China	CN	CHN	Guangdong	admin	26940000	1156237133
5	Mumbai	Mumbai	19.0761	72.8775	India	IN	IND	Mahārāshtra	admin	24973000	1356226629
6	Manila	Manila	14.5958	120.9772	Philippines	PH	PHL	Manila	primary	24922000	1608618140
7	Shanghai	Shanghai	31.2286	121.4747	China	CN	CHN	Shanghai	admin	24073000	1156073548
8	São Paulo	Sao Paulo	-23.5504	-46.6339	Brazil	BR	BRA	São Paulo	admin	23086000	1076532519

```
!head -n2 ./worldcities.csv
```

```
"city","city_ascii","lat","lng","country","iso2","iso3","admin_name","capital","population","id"  
"Tokyo","Tokyo","35.6870","139.7495","Japan","JP","JPN","Tōkyō","primary","37785000","1392685764"
```

Comenzamos analizando la estructura de los datos para crear manualmente el esquema.

### World cities database

Accurate and up-to-date database of the world's cities and towns



<https://www.kaggle.com/datasets/juanmah/world-cities>

Una vez comprendidos los datos, creamos el esquema:

```
from pyspark.sql.types import StructType, StructField, StringType, DoubleType, LongType

schema_worldcities = StructType([
    StructField("city", StringType(), True),
    StructField("city_ascii", StringType(), True),
    StructField("lat", DoubleType(), True),
    StructField("lng", DoubleType(), True),
    StructField("country", StringType(), True),
    StructField("iso2", StringType(), True),
    StructField("iso3", StringType(), True),
    StructField("admin_name", StringType(), True),
    StructField("capital", StringType(), True),
    StructField("population", LongType(), True),
    StructField("id", LongType(), True)
])
```

Y procedemos a cargar los datos:

```
df_cities = ( spark.read  
              .format("csv")  
              .schema(schema_worldcities)  
              .option("header", "true")  
              .option("quote", "\"")  
              .load("./worldcities.csv")  
 )  
  
df_cities.printSchema()  
df_cities.show(5)
```

Como el fichero tiene encabezado se lo indicamos

En el fichero todos los campos están rodeados de comillas dobles. Con quote se lo indicamos de forma que las omita y procese únicamente lo que hay entre comillas.

```
"city","city_ascii","lat",  
"Tokyo","Tokyo","35.6870",
```

Con show(5) muestro las 5 primeras filas del *dataframe*

Comprobamos que el esquema del *dataframe* se corresponde con lo que le indicamos

```
root
|-- city: string (nullable = true)
|-- city_ascii: string (nullable = true)
|-- lat: double (nullable = true)
|-- lng: double (nullable = true)
|-- country: string (nullable = true)
|-- iso2: string (nullable = true)
|-- iso3: string (nullable = true)
|-- admin_name: string (nullable = true)
|-- capital: string (nullable = true)
|-- population: long (nullable = true)
|-- id: long (nullable = true)
```

Y podemos ver los datos ya cargados

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| city|city_ascii|    lat|     lng| country|iso2|iso3| admin_name|capital|population|      id|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Tokyo|    Tokyo| 35.687|139.7495|   Japan| JP| JPN|    Tōkyō|primary| 37785000|1392685764|
| Jakarta|  Jakarta| -6.175|106.8275|Indonesia| ID| IDN|    Jakarta|primary| 33756000|1360771077|
| Delhi|    Delhi|  28.61|   77.23|    India| IN| IND|    Delhi| admin| 32226000|1356872604|
| Guangzhou| Guangzhou| 23.13| 113.26|    China| CN| CHN|    Guangdong| admin| 26940000|1156237133|
| Mumbai|   Mumbai|19.0761| 72.8775|    India| IN| IND|Mahārāshtra| admin| 24973000|1356226629|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

## Carga de datos JSON

Vamos a ver cómo cargar datos de un JSON, en esta ocasión del *dataset* del Titanic



ABDUL BASIT AI · UPDATED 3 MONTHS AGO

### Titanic-json-format

Titanic dataset in JSON format — ready to use with Pandas and APIs.

1 Code Download



<https://www.kaggle.com/datasets/engrbasit62/titanic-json-format>

```
titanic.json X dataframes_sample  
root 891 items  
0  
PassengerId "1"  
Survived "0"  
Pclass "3"  
Name ""Braund, Mr. Owen Harris""  
Sex "male"  
Age "22"  
SibSp "1"  
Parch "0"  
Ticket "A/5 21171"  
Fare "7.25"  
Cabin ""  
Embarked "S"
```

...

## Comenzamos creando el esquema

```
# Definición del esquema para el dataset Titanic
schema_titanic = StructType([
    StructField("PassengerId", IntegerType(), True),
    StructField("Survived", IntegerType(), True),      # 0 o 1
    StructField("Pclass", IntegerType(), True),        # 1, 2 o 3
    StructField("Name", StringType(), True),
    StructField("Sex", StringType(), True),
    StructField("Age", DoubleType(), True),             # Double porque hay edades fraccionarias (ej: 0.42)
    StructField("SibSp", IntegerType(), True),          # Esposos/Hermanos a bordo
    StructField("Parch", IntegerType(), True),           # Padres/Hijos a bordo
    StructField("Ticket", StringType(), True),
    StructField("Fare", DoubleType(), True),             # Tarifa (dinero)
    StructField("Cabin", StringType(), True),
    StructField("Embarked", StringType(), True)          # Puerto de embarque (C, Q, S)
])
```

Y cargamos los datos

```
df_titanic = ( spark.read  
                .format("json")  
                .schema(schema_titanic)  
                .option("multiline", "true")  
                .load("titanic.json")  
            )
```

```
df_titanic.printSchema()  
df_titanic.show(5)
```

Le tenemos que avisar a Spark de que el fichero no es JSONL y tiene varias líneas por registro

Por defecto, Spark espera el formato **JSONL** (JSON Lines) donde:

- Cada línea es un objeto JSON completo y válido
- No hay corchetes envolviendo todo el archivo
- No hay comas al final separando los objetos

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
NULL	NULL	NULL	"Braund, Mr. Owen...	male	NULL	NULL	NULL	A/5 21171	NULL		S
NULL	NULL	NULL	"Cumings, Mrs. Jo...	female	NULL	NULL	NULL	PC 17599	NULL	C85	C
NULL	NULL	NULL	"Heikkinen, Miss....	female	NULL	NULL	NULL	STON/O2. 3101282	NULL		S
NULL	NULL	NULL	"Futrelle, Mrs. J....	female	NULL	NULL	NULL	113803	NULL	C123	S
NULL	NULL	NULL	"Allen, Mr. Willi...	male	NULL	NULL	NULL	373450	NULL		S

Cuando leemos ficheros puede que encontremos **datos corruptos o malformados**.

Debemos indicarle a Spark como actuar ante estos datos mediante  
.option("mode", "mode\_name")

Los modos que hay son:

- **PERMISSIVE:**

- Es el modo por defecto
- Si hay un dato que no encaja con el tipo esperado **no falla ni elimina la fila**
- Convierte el valor corrupto a `null` y carga el resto de la fila
- El problema es que se puede llenar el DataFrame de nulos silenciosamente y estropear cálculos sin darnos cuenta.
- En este modo podemos configurar una columna especial llamada `_corrupt_record` para que Spark guarde ahí el dato original sucio y se pueda revisar luego.

## Ejemplo de cómo guardar los datos corruptos en una columna aparte

```
schema = StructType([
    StructField("id", IntegerType(), True),
    StructField("nombre", StringType(), True),
    StructField("edad", IntegerType(), True),
    StructField("_corrupt_record", StringType(), True)
])

df = ( spark.read
        .format("csv")
        .option("header", "true")
        .schema(schema)
        .option("mode", "PERMISSIVE")
        .option("columnNameOfCorruptRecord", "_corrupt_record")
        .load("datos_sucios.csv")
)
```

En esta columna guardaremos los datos corruptos

Debemos usar la opción columnNameOfCorrruptRecord

- **DROPMALFORMED:**

- Si una fila tiene datos corruptos **la ignora por completo**
- Con esto aseguramos que los datos estén limpios, pero en un *dataset* con muchos datos mal formados perderemos muchas líneas y Spark no nos avisará explícitamente.

- **FAILFAST:**

- Al primer error que encuentre, Spark **detiene todo el proceso** inmediatamente y lanza una excepción
- Garantiza al 100% la calidad de los datos
- Se usa en sistemas críticos (p.e. transacciones bancarias) donde no se puede permitir ni un error o dato sucio.

# 2

## 2.3

# DATAFRAMES: INGESTA Y TRANSFORMACIÓN

## OPERACIONES BÁSICAS

Una vez cargados los datos en el DataFrame, lo habitual es prepararlos antes del análisis seleccionando columnas, filtrando filas, corrigiendo valores y creando nuevas métricas.

Es importante recordar que los DataFrames son **inmutables**, es decir, las operaciones no modifican el original, sino que devuelven un nuevo DataFrame con los cambios aplicados.

## SELECCIÓN: SELECT

Equivale al SELECT de SQL y sirve para **seleccionar una serie de columnas** y descartar el resto.

```
from pyspark.sql import functions as F
from pyspark.sql.functions import col, lit

df_cities.select("city", "population").show(3)

+-----+-----+
| city|population|
+-----+-----+
| Tokyo| 37785000|
| Jakarta| 33756000|
| Delhi| 32226000|
+-----+-----+
only showing top 3 rows
```

Por convención, las funciones de PySpark se importan como F

Indico las columnas con las que me quiero separar separadas por comas

En lugar de indicar el nombre de la columna, puedo usar la función `col()` que añade más funcionalidades

```
df_cities.select(col("city"), col("population")/1000000).show(3)
```

```
+-----+  
| city|(population / 1000000)|  
+-----+  
| Tokyo| 37.785 |  
| Jakarta| 33.756 |  
| Delhi| 32.226 |  
+-----+  
only showing top 3 rows
```

Puedo realizar **operaciones** con todos los elementos de la columna. En este ejemplo, mostrar los habitantes por millones

```
df_cities.select(col("city"), col("population").alias("población")).show(3)
```

```
+-----+  
| city|población|  
+-----+  
| Tokyo| 37785000 |  
| Jakarta| 33756000 |  
| Delhi| 32226000 |  
+-----+  
only showing top 3 rows
```

Puedo aplicar la función **alias()** para **cambiar el nombre** de la columna

```
df_cities.select(col("city"), col("population").cast("float")).show(3)
```

```
+-----+  
| city|population|  
+-----+  
| Tokyo| 3.7785E7|  
| Jakarta| 3.3756E7|  
| Delhi| 3.2226E7|  
+-----+  
only showing top 3 rows
```

Otra opción es usar la función `cast()` para **cambiar el tipo de datos** de la columna

```
df_cities.select(col("city"), col("population")>10000000).show(3)
```

```
+-----+  
| city|(population > 10000000)|  
+-----+  
| Tokyo| true|  
| Jakarta| true|  
| Delhi| true|  
+-----+  
only showing top 3 rows
```

También podemos **generar una columna True/False** según el valor correspondiente cumpla una condición

## RENOMBRADO: `withColumnRenamed()`

Renombra una columna del DataFrame

```
df_cities.withColumnRenamed("Population", "Población").show(3)
```

```
+-----+-----+-----+-----+-----+-----+-----+
| city|city_ascii|  lat|    lng| country|iso2|iso3|admin_name|capital|Población|
+-----+-----+-----+-----+-----+-----+-----+
| Tokyo|    Tokyo|35.687|139.7495|   Japan| JP| JPN|      Tōkyō|primary| 37785000|13926|
| Jakarta|  Jakarta|-6.175|106.8275|Indonesia| ID| IDN|     Jakarta|primary| 33756000|13607|
| Delhi|    Delhi| 28.61|   77.23|   India| IN| IND|      Delhi| admin| 32226000|13568|
+-----+-----+-----+-----+-----+-----+-----+
only showing top 3 rows
```

## FILTRADO DE FILAS: `where()` y `filter()`

Ambas funciones son sinónimos exactos y se pueden usar indistintamente.

**Filtran los registros** del DataFrame según un criterio.

Los operadores Python no funcionan, sino que se deben usar los operadores `&` (AND), `|` (OR) y `~` (NOT)

Cuando hay condiciones compuestas es obligatorio el uso de paréntesis

```
df_cities.filter( col("country") == "Spain" ).show(3)
```

```
+-----+-----+-----+-----+-----+-----+-----+
|    city|city_ascii|      lat|      long|country|iso2|iso3|admin_name|capital|population|
+-----+-----+-----+-----+-----+-----+-----+
| Madrid|    Madrid|40.4169|-3.7033| Spain|   ES| ESP|    Madrid|primary| 6211000|17246
| Barcelona| Barcelona|41.3833| 2.1833| Spain|   ES| ESP| Catalonia| admin| 4800000|17245
| Valencia| Valencia| 39.47|-0.3764| Spain|   ES| ESP| Valencia| admin| 1595000|17245
+-----+-----+-----+-----+-----+-----+-----+
only showing top 3 rows
```

Ejemplo de condición compuesta

```
(df_cities
    .filter( col("country") == "Spain" & col("population") > 5000000 )
    .select( "city", "country", "population" )
    .show(3)
)
```

```
+---+---+-----+
| city|country|population|
+---+---+-----+
|Madrid| Spain| 6211000|
+---+---+-----+
```

## CREACIÓN DE COLUMNAS: `withColumn()`

Este método sirve para dos cosas:

- Crear una columna nueva (si el nombre no existe)
- Sobrescribir una columna existente (si el nombre existe)

```
(df_cities
    .withColumn("million_pop", col("population")/1000000)
    .select("city", "population", "million_pop")
    .show(3)
)
```

```
+-----+-----+
| city|population|million_pop|
+-----+-----+
| Tokyo| 37785000|      37.785|
| Jakarta| 33756000|      33.756|
| Delhi| 32226000|      32.226|
+-----+
only showing top 3 rows
```

Si queremos que todas las columnas tengan un valor constante tenemos que utilizar la función lit()

```
(df_cities
    .withColumn("year", lit(2025))
    .select("city", "population", "year")
    .show(3)
)
```

```
+-----+-----+
| city|population|year|
+-----+-----+
| Tokyo| 37785000|2025|
| Jakarta| 33756000|2025|
| Delhi| 32226000|2025|
+-----+
only showing top 3 rows
```

La función lit la importamos de  
pyspark.sql.functions

Podemos realizar diferentes tipos de operaciones en `withColumn()`

## 1. Matemáticas y redondeo

Para columnas numéricas la clase F dispone de funciones para:

- **Redondeo:** `ceil` (techo), `floor` (suelo), `round` (redondeo), `bround` (redondeo bancario al par más cercano para reducir el sesgo en sumas grandes).
- **Logaritmos y exponentiales:** `log` (logaritmo natural), `log10` (logaritmo base 10), `exp` (exponencial  $e^x$ ), `factorial` (factorial de un entero).
- **Comparación entre columnas:** `max` y `min` (máximo / mínimo de una columna verticalmente), `greatest(col1, col2, ...)` y `least (col1, col2, ...)` (máximo / mínimo entre varias columnas de una misma fila)
- **Trigonometría:** `sin`, `cos`, `tan`, `asin`, `acos`, `atan`, `degrees` (radianes a grados), `radians` (grados a radianes)
- **Aleatoriedad:** `rand` (número entre 0 y 1), `randn` (número aleatorio de distribución normal o gausiana)

**Ejemplo:** redondear latitud y longitud a 2 decimales

Country	City	AccentCity	Region	Population	Latitude	Longitude
ad andorra la vella	Andorra la Vella	07	20430.0	42.5	1.5166667	
ad canillo	Canillo	02	3292.0	42.5666667	1.6	
ad encamp	Encamp	03	11224.0	42.533333299999995	1.5833333	
ad la massana	La Massana	04	7211.0	42.55	1.5166667	
ad les escaldes	Les Escaldes	08	15854.0	42.5	1.5333333	

```
from pyspark.sql.functions import round, col

df_redondeado = ( df_cities.withColumn("Latitude", round(col("Latitude"), 2))
                  .withColumn("Longitude", round(col("Longitude"), 2))
                )

df_redondeado.show(5)
```

Country	City	AccentCity	Region	Population	Latitude	Longitude
ad andorra la vella	Andorra la Vella	07	20430.0	42.5	1.52	
ad canillo	Canillo	02	3292.0	42.57	1.6	
ad encamp	Encamp	03	11224.0	42.53	1.58	

## 2. Manipulación de fechas y tiempo

Tenemos las siguientes funciones para fechas:

- `to_date / to_timestamp`: convertir strings a fechas
- `datediff`: restar fechas (devuelve número de días)
- `date_add / date_sub`: sumar/restar días a una fecha
- `year, month, dayofmonth`: extraer partes de una fecha

### 3. Manipulación de cadenas

Para el trabajo con cadenas hay un montón de funciones:

- **Limpieza y formato:**
  - **trim()**: elimina espacios blancos a ambos lados
  - **ltrim() / rtrim()**: elimina espacios solo a izquierda o derecha
  - **upper() / lower()**: convierte a mayúsculas/minúsculas
  - **initcap()**: tipo título
- **Unión y división:**
  - **concat(col1, col2, ...)**: une columnas sin separador. Si alguna columna es null, el resultado será null.
  - **concat\_ws(sep, col1, col2, ...)**: concatena columnas usando un separador. Si hay nulos se los salta.
  - **split(col, sep)**: divide un string en una lista

**Ejemplo:** añadir una columna eliminando espacios en blanco antes y después

```
from pyspark.sql.functions import col, trim

df_cities = df_cities.withColumn("City_Clean", trim(col("City")))
df_cities.show(3)
```

```
+-----+-----+-----+-----+-----+-----+
|Country|      City| AccentCity|Region|Population|      Latitude|Longitude|    City_Clean|
+-----+-----+-----+-----+-----+-----+
|  ad|andorra la vella|Andorra la Vella|    07|   20430.0|        42.5|1.51666667|andorra la vella|
|  ad|       canillo|      Canillo|    02|    3292.0|        42.5666667|      1.6|       canillo|
|  ad|       encamp|      Encamp|    03|  11224.0|42.533333299999995|1.5833333|       encamp|
+-----+-----+-----+-----+-----+-----+
only showing top 3 rows
```

Ejemplo: crear una nueva columna que será la concatenación de latitud y longitud

```
from pyspark.sql.functions import col, concat, lit

df_cities = df_cities.withColumn("Coordinates",
    concat(
        round( col("Latitude"), 2 ).cast("string"),
        lit(", "),
        round( col("Longitude"), 2).cast("string")
    )
)

df_cities.select("City", "Latitude", "Longitude", "Coordinates").show(5)
```

City	Latitude	Longitude	Coordinates
andorra la vella	42.5	1.5166667	42.5, 1.52
canillo	42.5666667	1.6	42.57, 1.6
encamp	42.533333299999995	1.5833333	42.53, 1.58

- **Extracción y longitud:**
  - **length()**: devuelve el número de caracteres
  - **substring(col, pos, len)**: extrae una parte del texto. *Ojo: los índices comienzan por 1.*
  - **instr(col, sub)**: busca la posición de una subcadena (0 si no la encuentra).
- **Reemplazo y expresiones regulares:**
  - **translate(col, "ABC", "123")**: reemplazo carácter a carácter (A->1, B->2, ..) Muy rápido para limpiezas simples
  - **regexp\_replace(col, pattern, replacement)**: busca un patrón regexp y lo reemplaza
  - **regexp\_extract(col, pattern, group\_idx)**: extrae algo específico que cumpla un patrón (p.e. el dominio de un email)
- **Relleno:**
  - **lpad / rpad(col, len, pad)**: rellena por la izquierda o la derecha

Ejemplo: queremos detectar países cuyo código excede de 2 caracteres

```
from pyspark.sql.functions import col, length

df_errores = df_cities.filter(length(col("Country")) != 2)

print("Registros con códigos de país sospechosos:")
df_errores.select("Country", "City").show()
```

Registros con códigos de país sospechosos:

```
[Stage 11:>
+----+----+
|Country|City|
+----+----+
+----+----+
```

#### 4. Manejo de nulos

Para el manejo de nulos Spark dispone de la función **coalesce()**.

Esta función se puede ver como una cascada de opciones que recibe varias columnas o valores y **devuelve el primero que no sea nulo**.

Permite fusionar información dispersa en varias columnas para crear una columna maestra

```
df = df.withColumn("email_final",
    coalesce(
        col("email_personal"),
        col("email_trabajo"),
        lit("No disponible")
    )
)
```

Ejemplo: quiero hacer una campaña de envío de correos, pero de algunos usuarios tengo solo el personal, de otros solo el del trabajo y de otros ninguno.

### 3. Tipos complejos

PySpark puede manejar **estructuras de datos** dentro de una sola celda.

Para convertir datos planos en lisas tenemos las siguientes funciones:

- **split(col, sep)**: convierte una cadena de texto en un array cortando por el delimitador.
- **array(col1, col2, ...)**: junta los valores de varias columnas del mismo registro en una sola lista.

```
( df_titanic
    .withColumn('sep_name', F.split(col("name"), " "))
    .select("sep_name", "Name")
    .show(3, truncate=False)
)
```

sep_name	Name
["Braund,, Mr., Owen, Harris"]	"Braund, Mr. Owen Harris"
["Cumings,, Mrs., John, Bradley, (Florence, Briggs, Thayer)"]	"Cumings, Mrs. John Bradley"
["Heikkinen,, Miss., Laina"]	"Heikkinen, Miss. Laina"

Una vez que tenemos una lista o array, la pregunta es ¿qué podemos hacer con él?

- **size(col)**: devuelve el número de elementos (longitud)
- **array\_contains(col, value)**: devuelve True si el valor está en la lista
- **col[i]**: permite acceder a un elemento específico (ojo, aquí el primer elemento es el índice 0)

```
# 1. Filtrar productos que tengan la etiqueta "urgente"
df.filter(F.array_contains(col("array_etiquetas"), "urgente"))

# 2. Crear columna con el primer elemento de la lista
df = df.withColumn("etiqueta_principal", col("array_etiquetas")[0])

# 3. Filtrar listas vacías
df.filter(F.size(col("array_etiquetas")) > 0)
```

La otra opción es **descomponer el array** con la función **explode()**.

Esta función toma un array en una fila y **genera una nueva fila por cada elemento** del array.

```
from pyspark.sql.functions import explode

# Supongamos df:
# +-----+
# | id/ array_etiquetas|
# +-----+
# | 1/ ["cine", "cena"] |
# +-----+  
  
df_explotado = df.withColumn("etiqueta_individual", explode(col("array_etiquetas")))

# Resultado df_explotado:
# +-----+-----+
# | id/ array_etiquetas/etiqueta_individual|  
# +-----+-----+-----+-----+-----+  
# | 1/ ["cine", "cena"] /                cine/  
# | 1/ ["cine", "cena"] /                cena/  
# +-----+-----+-----+-----+
```

Ojo, si tenemos listas con muchos elementos el DataFrame crecerá exponencialmente

Si tenemos varias columnas con arrays, podremos realizar operaciones entre ellos tratándolos como conjuntos:

- **array\_intersect**(col1, col2): elementos comunes
- **array\_union**(col1, col2): unión sin duplicados
- **array\_except**(col1, col2): resta de conjuntos (elementos que están en col1 y no en col2)
- **array\_distinct**(col): elimina duplicados dentro de un mismo array

---

```
# Escenario: Comparar respuestas correctas vs respuestas del alumno
# col("clave"): ["A", "B", "C"]
# col("alumno"): ["A", "C", "D"]

# ¿Cuáles acertó?
df = df.withColumn("aciertos", F.array_intersect(col("clave"), col("alumno")))
# Resultado: ["A", "C"]
```

Por último, también podemos convertir el array en una cadena simple, por ejemplo, si queremos exportar a CSV el DataFrame.

- **array\_join(col, delimiter)**: convierte lista a texto

```
df = df.withColumn("texto_final", F.array_join(col("array_etiquetas"), " - "))
```

# 2

## 2.4

# DATAFRAMES: INGESTA Y TRANSFORMACIÓN

## AGREGACIONES Y AGRUPAMIENTOS

Una parte muy importante cuando trabajamos con datos son las agregaciones, donde agrupamos una serie de registros según un campo para realizar algún tipo de operación.

**Ejemplo:** Calcular el promedio de ventas por provincia

## AGRUPAMIENTO BÁSICO (groupBy)

La forma más sencilla de agrupar es mediante `groupBy`.

Devuelve un objeto de tipo `RelationalGroupedDataset`, sobre el cual debemos llamar a una función de agregación

```
df.groupBy("categoria").count().show()
```

## AGREGACIONES MÚLTIPLES (agg)

Para realizar múltiples agregaciones, personalizar nombres o usar funciones complejas, **siempre** debemos usar .agg.

```
resultado = df.groupBy("tienda").agg(  
    F.sum("ingresos").alias("total_ventas"), # Suma total  
    F.avg("ingresos").alias("ticket_promedio"), # Media  
    F.max("unidades").alias("max_unidades_vendidas"), # Máximo  
    F.count("producto").alias("num_transacciones") # Conteo )  
resultado.show()
```

Es importante usar .alias ya que, en caso contrario, Spark pondrá los nombres a las columnas

## AGREGACIONES MÚLTIPLES (agg)

Para realizar múltiples agregaciones, personalizar nombres o usar funciones complejas, **siempre** debemos usar .agg.

```
resultado = ( df.groupBy("tienda")
    .agg( F.sum("ingresos").alias("total_ventas"),
          F.avg("ingresos").alias("ticket_promedio"),
          F.max("unidades").alias("max_unidades_vendidas"),
          F.count("producto").alias("num_transacciones"))
    )
resultado.show()
```

Es importante usar .alias ya que, en caso contrario, Spark pondrá los nombres a las columnas

¿Qué tipo de **funciones de agregación** tenemos disponibles? Hay múltiples funciones que podemos clasificar en varias categorías:

## FUNCIONES BÁSICAS

`count(col)`

Cuenta filas **no nulas** de la columna. Si usas `count("*")` o `count(1)`, cuenta **todas** las filas (incluyendo nulos) y es más rápido.

```
( df.groupBy("Crop")
    .agg(
        F.count("*").alias("num_cultivos")
    )
).show()
```

```
+-----+
| Crop|num_cultivos|
+-----+
| ... |          |
+-----+
```

## countDistinct(col)

Cuenta valores únicos exactos. Es un procedimiento costoso (shuffle)

```
( df.groupBy("Crop")
    .agg(
        F.countDistinct("Region").alias("Regiones diferentes")
    )
).show()
```

Crop	Regiones diferentes
Maize	4
Wheat	4
Rice	4
Barley	4

```
sum(col)
```

Suma valores numéricos. Ignora nulos.

```
( df.groupBy("Region")
    .agg(
        F.countDistinct("Rainfall_mm").alias("Total lluvia")
    )
).show()
```

Region	Total lluvia
Region_D	2276
Region_C	2222
Region_A	2321
Region_B	2282

```
avg(col) / mean(col)
```

Calcula el promedio aritmético. Ignora nulos

```
( df.groupBy("Region")
    .agg(
        F.mean("Temperature_C").alias("Temp. media")
    )
).show()
```

```
+-----+-----+
| Region|      Temp. media|
+-----+-----+
|Region_D| 24.87680577849118|
|Region_C| 25.106295993458712|
|Region_A| 24.86243654822333|
|Region_B| 25.06269492203117|
+-----+-----+
```

```
min(col) / max(col)
```

Valores máximo y mínimo. Funciona con números, textos y fechas

```
( df.groupBy("Region")
    .agg(
        F.max("Yield_ton_per_ha").alias("Rendimiento máximo"),
        F.min("Yield_ton_per_ha").alias("Rendimiento mínimo")
    )
).show()
```

Region	Rendimiento máximo	Rendimiento mínimo
Region_D	207.21	31.17
Region_C	205.19	32.88
Region_A	203.29	30.14
Region_B	205.23	28.45

## FUNCIONES ESTADÍSTICAS PARA IA Y DATA SCIENCE

Estas funciones son críticas para **entender la distribución** de los datos antes de entrenar modelos (detección de outliers, normalización)

`stddev(col)`

La **desviación estándar** es una función fundamental para entender la **estabilidad y la distribución** de los datos.

Se representa como  $\sigma$  y básicamente mide la **dispersión** de los datos respecto a la media:

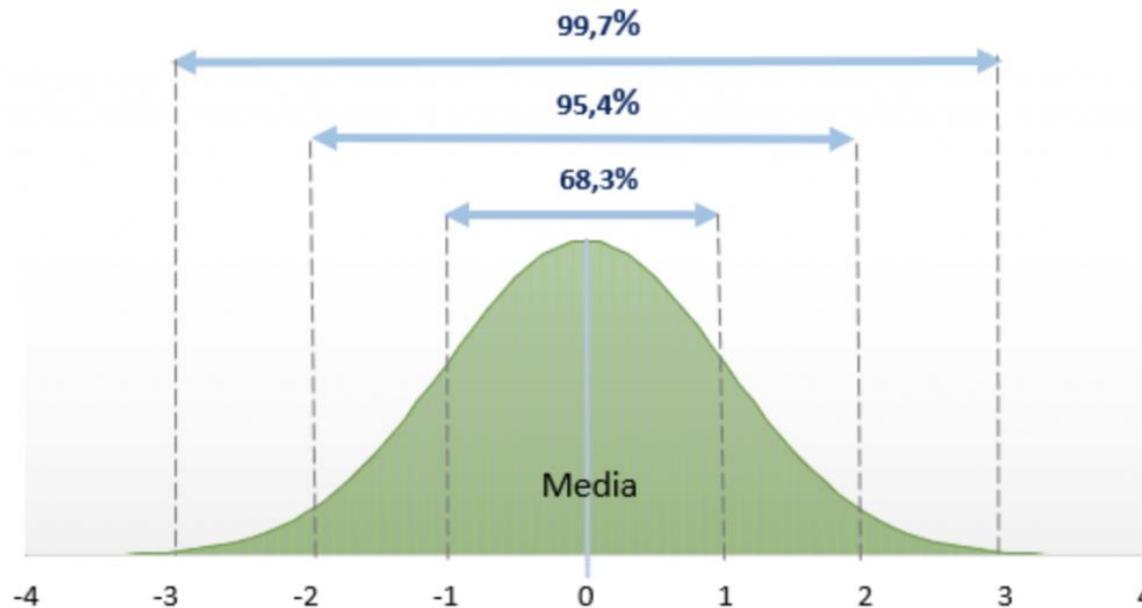
- Una **desviación baja** indica que los datos están muy agrupados cerca del promedio -> comportamiento predecible
- Una **desviación alta** indica que los datos están muy dispersos, hay mucha variabilidad o ruido -> comportamiento volátil.

La desviación estándar se utiliza habitualmente en la fase de **Feature Engineering**:

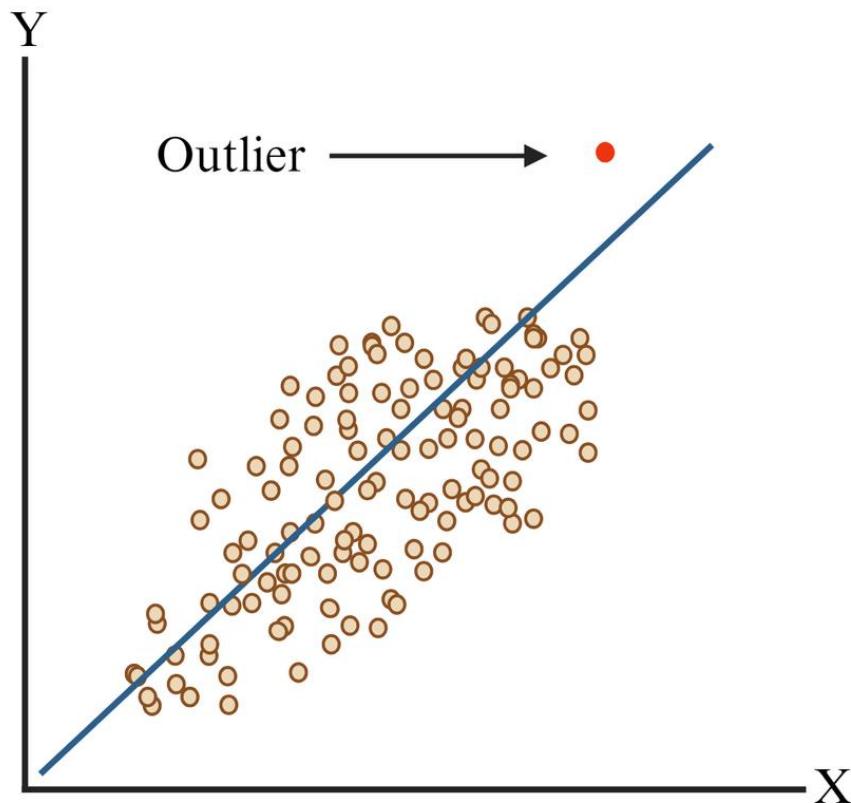
### Detección de outliers (Regla de las 3 sigmas)

En una distribución normal (Campana de Gauss):

- El **68%** de los datos está a  $\pm 1$  desviación estándar de la media
- El **95%** de los datos está a  $\pm 2$
- El **99.7%** de los datos está a  $\pm 3$



Si un dato está a más de 3 veces la desviación estándar de distancia de la media, es un **outlier (anomalía)** y probablemente deba ser eliminado o investigado antes de entrenar el modelo.



## Estandarización (Z-Score Scaling)

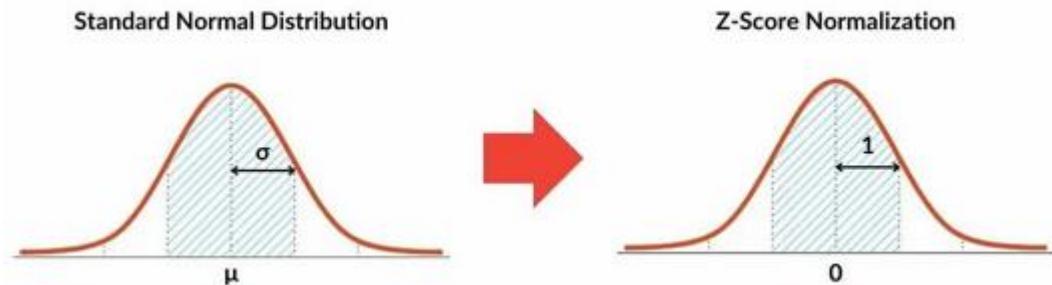
Muchos algoritmos de IA (redes neuronales, K-Means, SVM) fallan si las variables tienen escalas distintas (p.e. edad de 0-100 y salario de 0-1000000)

Para solucionarlo, se deben transformar los datos a una escala común donde **la media es 0 y la desviación estándar es 1**.

La fórmula es:

$$Z = \frac{x - \mu}{\sigma}$$

Score      Mean  
                ↓  
                ↓  
 $\sigma$       SD



```
stats = ( df.groupBy("Crop")
            .agg(
                F.mean("Temperature_C").alias("media"),
                F.stddev("Temperature_C").alias("desviacion")
            )
        )

# Unimos Las estadísticas con Los datos originales
df_con_stats = df.join(stats, "Crop") ← Hago un join para que estos
# datos estén disponibles en el
# dataframe

# 3. Calculamos el Z-Score para detectar anomalías
# Z-Score = (Valor - Media) / Desviación
df_analisis = df_con_stats.withColumn(
    "z_score",
    (F.col("Temperature_C") - F.col("media")) / F.col("desviacion")
)

df_analisis.select("Crop", "Temperature_C", "media", "desviacion", "z_score")
```

Calculo la media y la desviación estándar para cada tipo de cultivo

Hago un join para que estos datos estén disponibles en el dataframe

Crop	Temperature_C	media	desviacion	z_score
Maize	30.0	24.91130399677028	5.809848453283212	0.8758741375352124
Maize	15.5	24.91130399677028	5.809848453283212	-1.6198880353672294
Maize	28.1	24.91130399677028	5.809848453283212	0.5488432321204099
Maize	32.9	24.91130399677028	5.809848453283212	1.3750265721157005
Maize	20.6	24.91130399677028	5.809848453283212	-0.7420682366222323
Maize	31.4	24.91130399677028	5.809848453283212	1.116844278367172
Maize	32.2	24.91130399677028	5.809848453283212	1.2545415016997212
Maize	32.0	24.91130399677028	5.809848453283212	1.2201171958665835

Este sería el valor que debería reemplazar a la temperatura si fuéramos a usar estos datos en un algoritmo de ML

```
stats2 = ( df_analisis.groupBy("Crop")
            .agg(
                F.mean("z_score").alias("media"),
                F.stddev("z_score").alias("desviacion")
            )
)
stats2.show(truncate=False)
```

Crop	media	desviacion
Maize	1.3748478424445721E-15	0.9999999999999991
Wheat	4.6878564318142675E-15	0.9999999999999989
Rice	-7.48306078861112E-15	1.0000000000000002
Barley	6.627201342235676E-15	0.9999999999999981

Podemos comprobar que, efectivamente, la media de los nuevos datos es 0 y la desviación estándar 1

```
stddev_pop(col)
```

Cuenta filas **no nulas** de la columna. Si usas `count(*)` o `count(1)`, cuenta **todas** las filas (incluyendo nulos) y es más rápido.

```
variance(col)
```

La **varianza** ( $\sigma^2$ ) es el promedio de los cuadrados de las diferencias con la media.

$$\text{Varianza} = \sum_{I=1}^n \frac{(x_i - \bar{x})^2}{n}$$

Es una medida difícil de interpretar para humanos (estoy elevando al cuadrado, ¿qué significan, por ejemplo, grados cuadrados?) pero tiene mucho significado para las máquinas. Sus ventajas son:

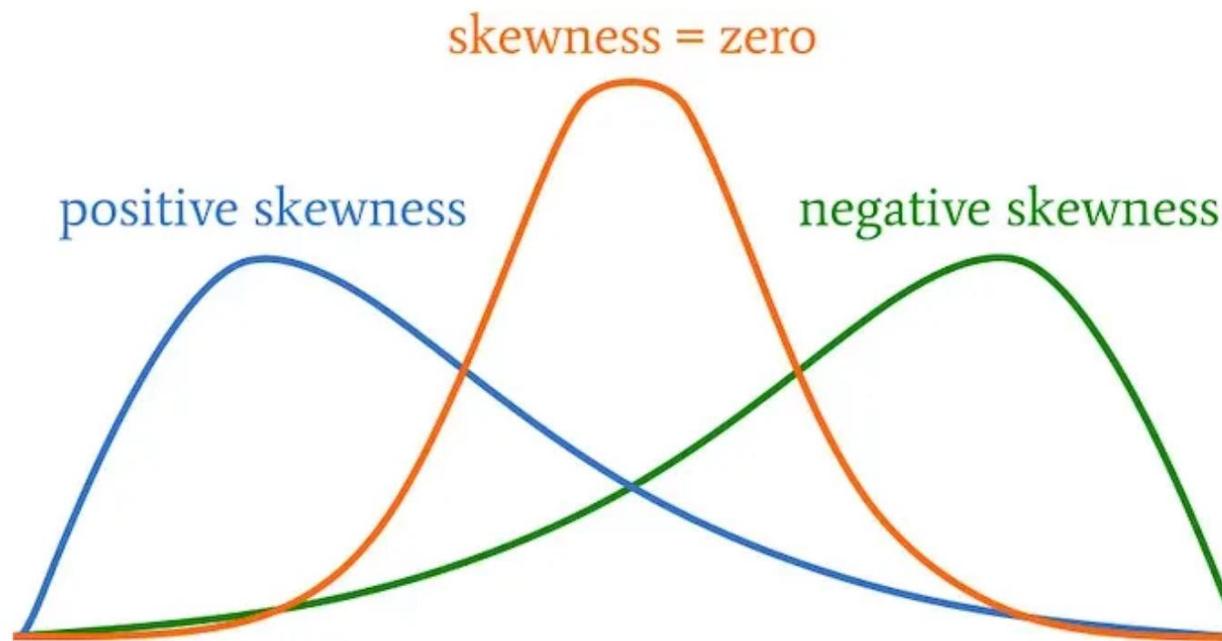
- **Eliminar negativos**: al elevar al cuadrado todos los valores son positivos
- **Castigar a los *outliers***, que destacan mucho más cuando se elevan al cuadrado.

La varianza se utiliza para:

- **Evaluación de los modelos:** si un modelo tiene varianza alta significa que ha memorizado el ruido de los datos de entrenamiento y que es inestable: si cambias un poco los datos de entrada, la predicción cambia radicalmente. Lo ideal es buscar **modelos con baja varianza (estables) y bajo sesgo (precisos)**
- **Selección de características:** si una columna tiene varianza 0 significa que todos los valores son iguales, por lo que no aporta información. Lo ideal es eliminar todas las columnas con varianzas cercanas a 0
- **PCA (Análisis de Componentes Principales):** cuando se quieren reducir dimensiones se usa PCA, que elige las columnas donde hay máxima varianza (la varianza es sinónimo de información).

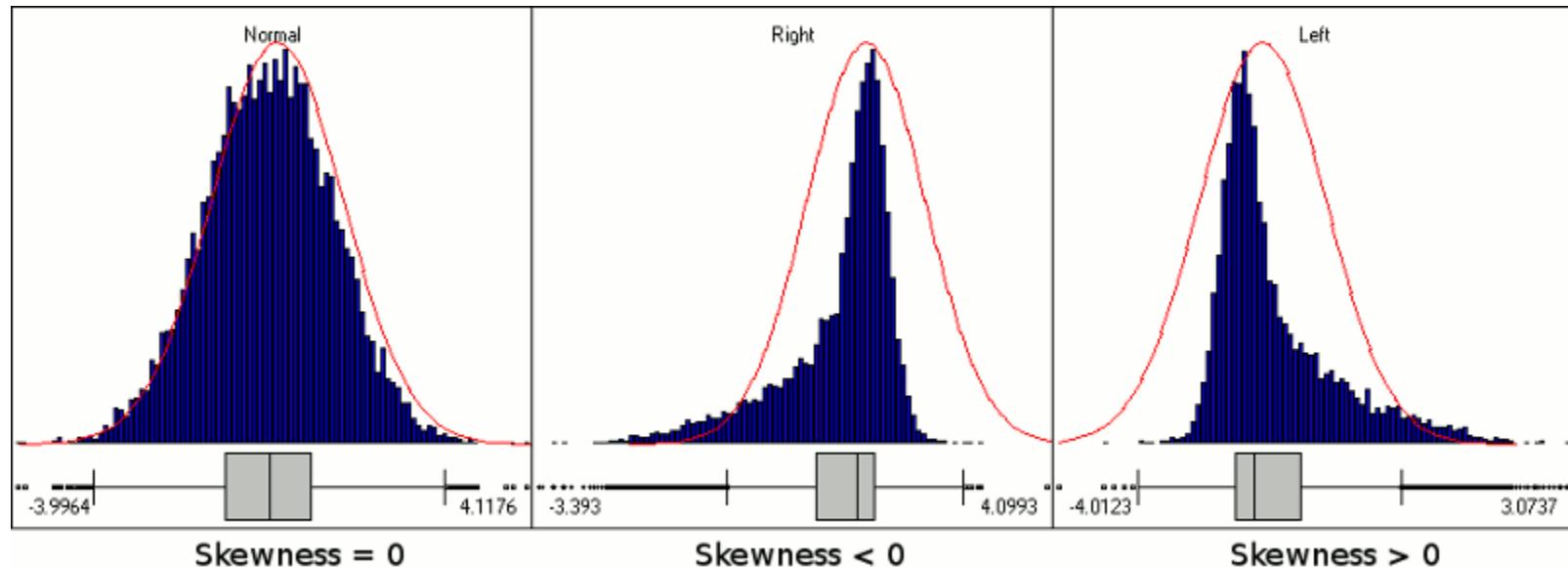
`skewness(col)`

Es la **asimetría** de la distribución. Si es  $>1$  o  $<-1$  la distribución de datos no es normal y se necesita realizar algún tipo de transformación



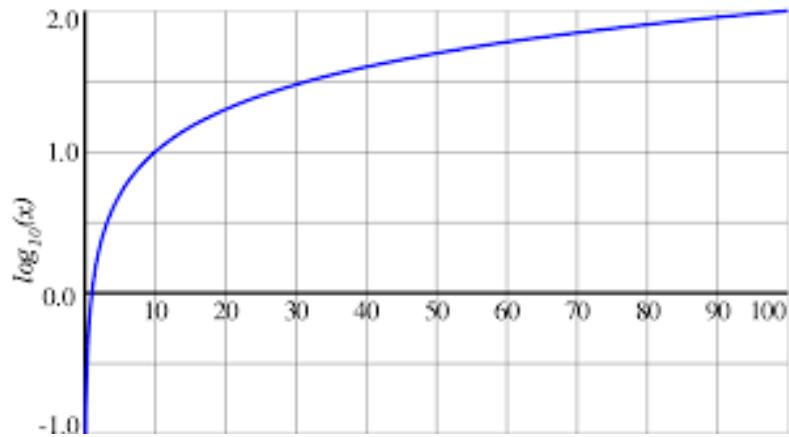
La mayoría de los modelos de ML asumen que los datos tienen forma de **campana perfecta** (distribución normal). Si la *skewness* es alta, el modelo aprenderá **sesgado**.

Lo ideal es que tenga un valor próximo a 0, donde la media y la mediana (*el valor que divide el dataset en dos mitades*) son iguales.



La forma de solucionar el *skewness* de los datos suele ser mediante la aplicación del **logaritmo** base 10 (recuerda que le logaritmo es la operación matemática inversa de la exponenciación).

El logaritmo reduce las distancias de forma **no lineal**. Cuanto más grande es el número, más fuerte lo castiga.



```
( df.groupBy("Crop")
    .agg(
        F.skewness("Temperature_C").alias("sesgo"),
    )
).show(truncate=False)
```

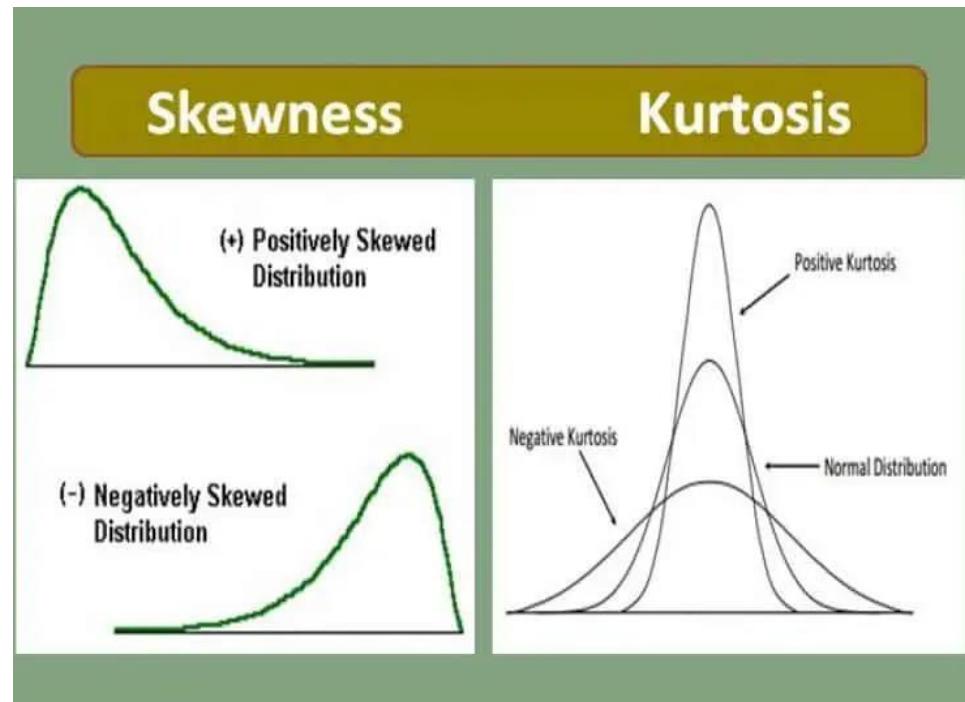
[Stage 87:>

Crop	sesgo
Maize	0.04038400226515987
Wheat	-0.023524078832991815
Rice	0.012614452271872583
Barley	0.009337968579507578

kurtosis(col)

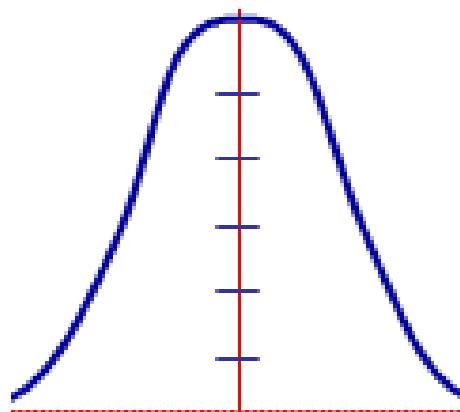
La **Curtosis** es el otro elemento (junto a la media, la desviación y el skewness) que debemos mirar.

Si el skewness nos dice hacia donde se inclina la montaña, la **curtosis** nos dice cuán gordas con las colas de esa montaña, ya que mide el *apuntamiento* y el grosor de las colas respecto a una distribución normal.

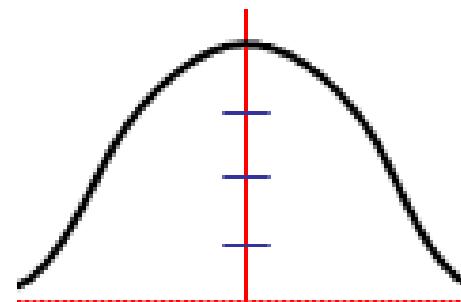


Los valores que podemos encontrar son:

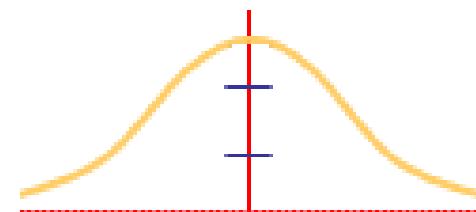
- **Valor próximo a 0 (mesocúrtica):**
  - Es el valor normal.
  - Los eventos ocurren con la frecuencia esperada. Ni muchas sorpresas ni pocas.
- **Valor superior a 0 (leptocúrtica):** pico alto y agudo en el centro, y colas largas y gordas en los extremos.



**Leptocúrtica**



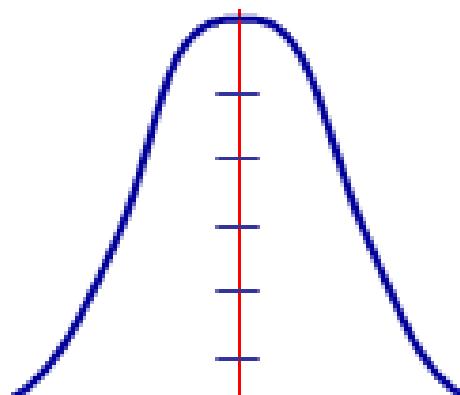
**Mesocúrtica**



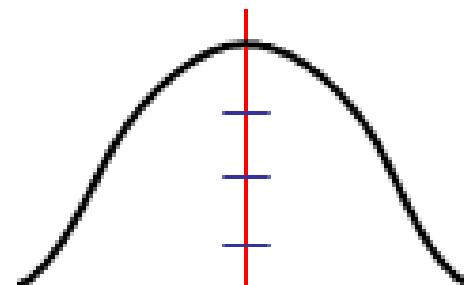
**Platicúrtica**

- **Valor superior a 0 (leptocúrtica):**

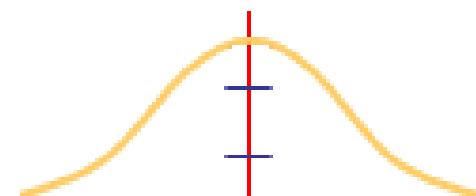
- Pico alto y agudo en el centro, y colas largas y gordas en los extremos.
- La mayoría de los datos están concentrados en el promedio. Pero hay probabilidad anormalmente alta de encontrar valores muy extremos  
**Ejemplo:** criptomonedas, la mayoría de los días no pasa nada, pero un día cae un 40%
- Es necesario limpiar outliers agresivamente



**Leptocúrtica**

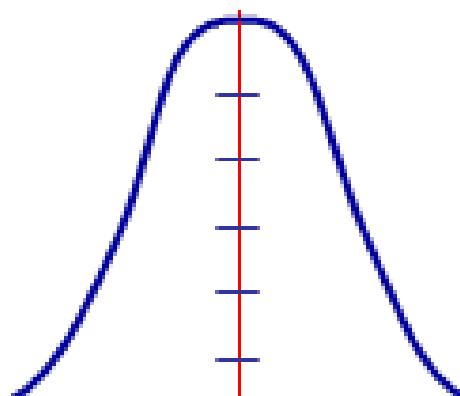


**Mesocúrtica**

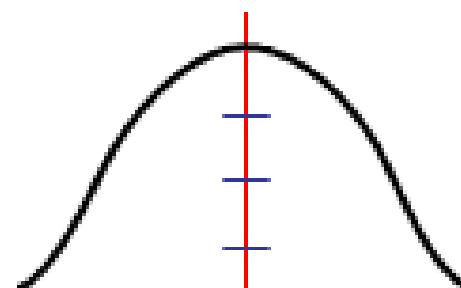


**Platicúrtica**

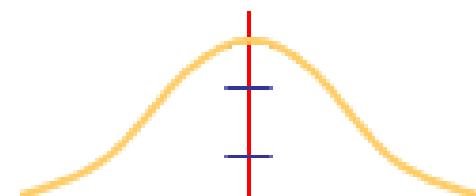
- **Valor inferior a 0 (plasticúrtica):**
  - La cima es plana y ancha, y las colas muy delgadas
  - Los datos están muy dispersos alrededor de la media
  - No hay valores extremos. Todo es zona gris
  - Los datos no tienen patrón claro. A los algoritmos de clustering (como K-means) les cuesta encontrar el centro



**Leptocúrtica**



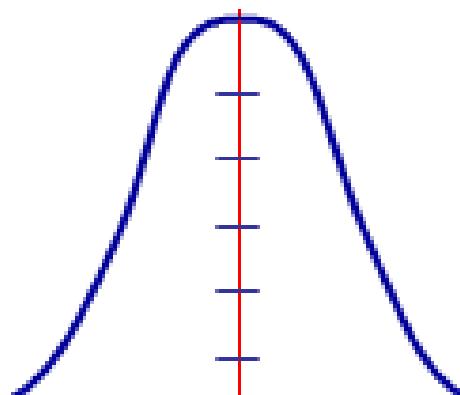
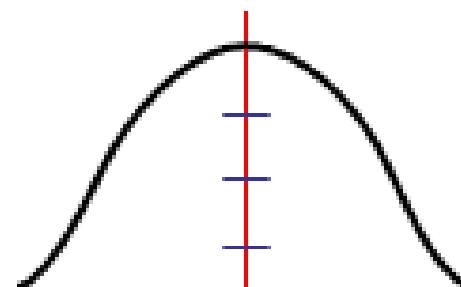
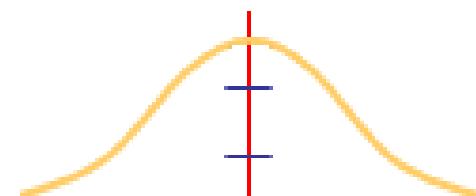
**Mesocúrtica**



**Platicúrtica**

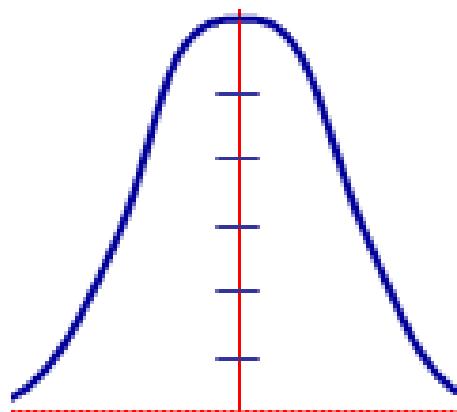
**Interpretación de los valores positivos de curtosis:**

- **0 a 1 (zona segura)**. La distribución es casi normal.
- **1 a 3 (Precaución)** Es claramente leptocúrtica. Hay más outliers de los que predice la campana de Gauss, pero suelen ser manejables
- **>3 (Curtosis alta)** La curtosis es muy alta. La probabilidad de eventos extremos (5 ó 6 desviaciones estándar) es muy alta
- **>10 (Curtosis extrema)**: suele indicar errores de medición

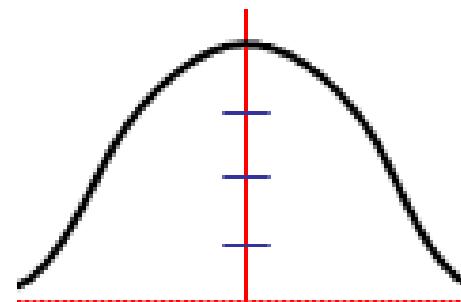
**Leptocúrtica****Mesocúrtica****Platicúrtica**

### Interpretación de los valores negativos de curtosis:

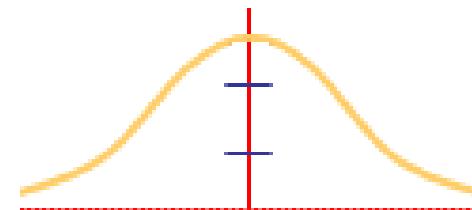
- **0 a -0.5 (casi normal).** Es una campana de Gauss un poco aplastada, pero sigue comportándose bien.
- **-0.5 a -1.0 (moderado).** La cima desaparece y empieza a parecer una colina suave. Los datos están dispersos, pero no extremos.
- **-1.2 (meseta).** Valor de referencia clásico que refleja una distribución uniforme.
- **-2.0 (binario).** Valor mínimo posible. Distribución de Bernoulli con  $p=0.5$  (p.e. lanzar moneda: 50% caras y 50% cruces)



**Leptocúrtica**



**Mesocúrtica**



**Platicúrtica**

```
( df.groupBy("Crop")
    .agg(
        F.kurtosis("Temperature_C").alias("curtosis"),
    )
).show(truncate=False)
```

Crop	curtosis
Maize	-1.20149109209531
Wheat	-1.1794781330751978
Rice	-1.2109153977374925
Barley	-1.2315894519670176

Observa que aquí los valores son todos próximos a 1.2, lo que denota una distribución uniforme. Casi seguro que debido a que este se trata de un **dataset sintético**.

## FUNCIONES DE APROXIMACIONES

Estas funciones son relevantes cuando la velocidad es más importante que la precisión exacta

`approx_count_distinct(col, rsd)`

Realiza un conteo único estimado utilizando HyperLogLog.

Mucho más rápido que `countDistinct(col)`. `rsd` es el error permitido (por defecto 0.05)

```
approx_percentile(col, rsd)
```

Calcula percentiles (mediana, Q1, Q3)

Calcular la mediana es muy costoso en distribuido. Esto lo hace viable.

## FUNCIONES DE SELECCIÓN POSICIONAL

Estas funciones son relevantes cuando la velocidad es más importante que la precisión exacta

```
first(col, ignorenulls=False)
```

x





`corr(col1, col2)`

Es la **asimetría** de la distribución. Si es  $>1$  o  $<-1$  la distribución de datos no es normal y se necesita realizar algún tipo de transformación

