In [2]:
```
#Dataset-1
from pyspark.sql import SparkSession
from pyspark.sql.types import *
spark = SparkSession.builder.getOrCreate()
store_df1 = spark.read.csv("USA.csv")

Schema=StructType([  StructField("ID",IntegerType(),nullable=True),
  StructField("Name",StringType(),nullable=True),
  StructField("VaccinationType",StringType(),nullable=True),
  StructField("VaccinationDate",StringType(),nullable=True),
StructField("Region",StringType(),nullable=False),
])


df1 = spark.read.option("header",True).schema(Schema).csv("USA.csv")
df1=df1.na.fill(value="USA",subset=["Region"])
df1.show()
```

```
+---+----+---------------+---------------+------+
| ID|Name|VaccinationType|VaccinationDate|Region|
+---+----+---------------+---------------+------+
|  1| Sam|            EFG|        6152022|   USA|
|  2|John|            XYZ|        1052022|   USA|
|  3|Mike|            ABC|       12282021|   USA|
+---+----+---------------+---------------+------+
```

In [6]:
```
#Dataset-2
from pyspark.sql.functions import col, to_date
store_df2 = spark.read.csv("AUS.csv")
Schema=StructType([  StructField("ID",IntegerType(),nullable=True),
  StructField("Name",StringType(),nullable=True),
  StructField("VaccinationType",StringType(),nullable=True),
StructField("DateOfBirth",StringType(),nullable=True),
  StructField("VaccinationDate",StringType(),nullable=True),
StructField("Region",StringType(),nullable=False),

])


df2 = spark.read.option("header",True).schema(Schema).csv("AUS.csv")
df2 = df2.withColumn('VaccinationDate',to_date(col('VaccinationDate'), 'dd-MM-yyyy
df2 = df2.withColumn('DateOfBirth',to_date(col('DateOfBirth'), 'dd-MM-yyyy'))
df2=df2.na.fill(value="AUS",subset=["Region"])
df2.show()
```

```
+---+---------+---------------+-----------+---------------+------+
| ID|     Name|VaccinationType|DateOfBirth|VaccinationDate|Region|
+---+---------+---------------+-----------+---------------+------+
|  1|     Mike|            LMN|       null|     2022-05-11|   AUS|
|  2|Jonnathan|            XYZ| 1997-12-13|           null|   AUS|
|  3| Cristina|            ABC| 1998-03-12|     2022-03-12|   AUS|
+---+---------+---------------+-----------+---------------+------+
```

In [7]:
```
#Dataset-3
store_df3 = spark.read.csv("IND.csv")

Schema=StructType([  StructField("ID",IntegerType(),nullable=True),
  StructField("Name",StringType(),nullable=True),
StructField("DateOfBirth",StringType(),nullable=True),
  StructField("VaccinationType",StringType(),nullable=True),
  StructField("VaccinationDate",StringType(),nullable=True),
```

```
StructField("Free/Paid",StringType(),nullable=True),
StructField("Region",StringType(),nullable=False),

])


df3 = spark.read.option("header",True).schema(Schema).csv("IND.csv")
df3 = df3.withColumn('VaccinationDate',to_date(col('VaccinationDate'), 'yyyy-MM-dd
df3 = df3.withColumn('DateOfBirth',to_date(col('DateOfBirth'), 'yyyy-MM-dd'))
df3=df3.na.fill(value="INDIA",subset=["Region"])
df3.show()
```

```
+---+------+-----------+---------------+---------------+---------+------+
| ID|  Name|DateOfBirth|VaccinationType|VaccinationDate|Free/Paid|Region|
+---+------+-----------+---------------+---------------+---------+------+
|  1| Vikas| 1998-12-01|            XYZ|     2022-01-01|        F| INDIA|
|  2| Rahul| 1982-08-13|            ABC|     2022-03-05|        P| INDIA|
|  3|Sameer| 1952-08-13|            ABC|     2022-02-20|        F| INDIA|
+---+------+-----------+---------------+---------------+---------+------+
```

In [8]:
```
#We are merging all DateFrames stored
import functools
def unionAll(dfs):
    return functools.reduce(lambda df1, df2: df1.union(df2.select(df1.columns)), d

unioned_df = unionAll([df1, df2, df3])
unioned_df.show()
```

```
+---+---------+---------------+---------------+------+
| ID|     Name|VaccinationType|VaccinationDate|Region|
+---+---------+---------------+---------------+------+
|  1|      Sam|            EFG|        6152022|   USA|
|  2|     John|            XYZ|        1052022|   USA|
|  3|     Mike|            ABC|       12282021|   USA|
|  1|     Mike|            LMN|     2022-05-11|   AUS|
|  2|Jonnathan|            XYZ|           null|   AUS|
|  3| Cristina|            ABC|     2022-03-12|   AUS|
|  1|    Vikas|            XYZ|     2022-01-01| INDIA|
|  2|    Rahul|            ABC|     2022-03-05| INDIA|
|  3|   Sameer|            ABC|     2022-02-20| INDIA|
+---+---------+---------------+---------------+------+
```

In [9]:
```
#To get Count of people by region
unioned_df.groupBy("Region").count().show(truncate=False)
```

```
+------+-----+
|Region|count|
+------+-----+
|USA   |3    |
|AUS   |3    |
|INDIA |3    |
+------+-----+
```

In [10]:
```
#To get Count of Vaccination Types
unioned_df.groupBy("VaccinationType").count().show(truncate=False)
```

```
+---------------+-----+
|VaccinationType|count|
+---------------+-----+
|EFG            |1    |
|XYZ            |3    |
|ABC            |4    |
|LMN            |1    |
+---------------+-----+
```

In [ ]: