### Methods to split a column in pyspark DataFrame

1. **Split a column which is a simple array**

**Input: an array of simple data type, e.g., [2,5]**

**Output: multiple columns, e.g., two columns, [2], [5]**

***- Using select()***

e.g., df\_crit\_E has a column saddle\_edge, which is an array of integer.

Graphical user interface, text, application

Description automatically generated

df\_crit\_E\_pts\_saddle = df\_crit\_E.select(col("Ver"), col("Saddle\_edge"), df\_crit\_E.Saddle\_edge[0], df\_crit\_E.Saddle\_edge[1])

df\_crit\_E\_pts\_saddle = df\_crit\_E\_pts\_saddle.select("Ver", col("Saddle\_edge[0]").alias("saddle\_pt1"), col("Saddle\_edge[1]").alias("saddle\_pt2"))

df\_crit\_E\_pts\_saddle.printSchema()

df\_crit\_E\_pts\_saddle.show()

1. **Flatten an array of array to one column**

**Input: an array of array, e.g., [[2,5],[5,8]]**

**Output: a column which is an array, may have duplicates. E.g., [2,5,5,8]**

**-Using flatten()**

df\_crit\_T\_LS\_flatten = df\_crit\_T\_LS.select('Ver', flatten('LS\_edge'))

df\_crit\_T\_LS\_flatten.printSchema()

# df\_crit\_T\_LS\_flatten.show()

**Notes: if we want to remove the duplicates inside an array, we can use the array\_distinct() function.**

from pyspark.sql.functions import array\_distinct

df\_crit\_T\_LS\_flatten\_dis = df\_crit\_T\_LS\_flatten.withColumn("arraycol\_without\_dupes", array\_distinct("flatten(LS\_edge)"))

1. **Split a column of nested struct to multiple columns**

**Input: a column of struct, e.g., [2,5]**

**Output: multiple columns, e.g., [2], [5], see an example below**

**-Using select()**

Table

Description automatically generated

df\_crit\_E\_pts\_saddle\_sel = df\_crit\_E\_pts\_saddle.select(col("Ver"), col("pts\_saddle.\*"))

# df2 = df\_crit\_E\_pts\_saddle\_sel.toDF("Ver", "pt1", "pt2") # this line is not necessary

df2.printSchema()

df2.show()

1. **Explode a row to multiple rows**

**Input: a column of array, e.g., [2,5,[3,4]]**

**Output: a sized number of rows, e.g., [2]; [5]; [3,4]**

**-Using explode()**

**Note: if the input column is an array of array, the output rows will be array.**

df3 = df\_crit\_VET.select("Ver", explode("Critical"))

df3.printSchema()

df3.show()