Exercice 01 : SELECT name FROM Customer WHERE month(startDate)=7

In [ ]:

```
import findspark
findspark.init()
import pyspark
```

In [ ]:

```
sc = pyspark.SparkContext(appName="exosql")
```

In [ ]:

```
rdd = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/Customer.txt")
rdd = rdd.map (lambda x : x.split(","))
rdd = rdd.map (lambda x : (x[1].split("/")[1],x[2]))
rdd = rdd.filter (lambda x : x[0] == "07" )
rdd = rdd.map (lambda x : x[1])
rdd.collect()
```

In [ ]:

```
rdd.count()
```

SELECT DISTINCT name FROM Customer WHERE month(startDate)=7

In [ ]:

```
rdd2 = rdd
rdd2.distinct().collect()
```

In [ ]:

```
rdd1 = rdd.map(lambda x : (x,1))
rdd1 = rdd1.reduceByKey(lambda x,y : x+y)
rdd1 = rdd1.map(lambda x : x[0])
rdd1.collect()
```

SELECT O.cid, SUM(total), COUNT(DISTINCT total) FROM Order O GROUP BY O.cid

In [ ]:

```
rdd = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/order.txt")
rdd = rdd.map (lambda x : x.split(","))
rdd = rdd.groupBy(lambda x : x[0])
rdd = rdd.map (lambda x : ( x[0], list(x[1])[0] ) )
rdd = rdd.map (lambda x : ( x[0], list (map (int, x[1])) ))
rdd = rdd.map (lambda x : ( x[0], sum(x[1]), len(set(x[1])) ))
rdd.collect()
```

SELECT C.cid, O.total FROM Customer C, Order O WHERE month(startDate)=7 and C.cid=O.cid

In [ ]:

```
rdd1 = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/Customer.txt")
rdd1 = rdd1.map (lambda x : x.split(","))

rdd2 = sc.textFile("C:/Users/EASYFRONT/Documents/BD/requetessql/data/Order.txt")
rdd2 = rdd2.map (lambda x : x.split(","))

rdd3 = rdd1.join(rdd2)
rdd3 = rdd3.map (lambda x : (x[0],(x[1][0].split("/")[1],x[1][1])))
rdd3 = rdd3.filter ( lambda x : x[1][0]=="07")
rdd3 = rdd3.map ( lambda x : (x[0],x[1][1]))
rdd3.collect()
```

In [ ]:

```
rdd3.count()
```

In [ ]: