from pyspark.sql import SparkSession

import getpass

username = getpass.getuser()

spark = SparkSession.builder. \

config('spark.ui.port', '0'). \

appName("Credit\_card"). \

config("spark.sql.warehouse.dir", f"/user/{username}/warehouse"). \

enableHiveSupport().master('yarn').getOrCreate()

from pyspark.sql.functions import col

from pyspark.sql.types import IntegerType, DoubleType, BooleanType, DateType

order\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/orders.csv')

account\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/account.csv')

card\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/card.csv')

clients\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/client.csv')

disp\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/disp.csv')

district\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/district.csv')

loan\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/loan.csv')

trans\_df = spark.read.format('csv').option('header', 'true').option('delimiter',';').option("mode", "DROPMALFORMED").load('/user/itv006617/data/input/trans.csv')

order\_df=order\_df.withColumn('order\_id', col('order\_id').cast(IntegerType())).withColumn('account\_id', col('account\_id').cast(IntegerType()))

account\_df=account\_df.withColumn('account\_id', col('account\_id').cast(IntegerType())).withColumn('district\_id', col('district\_id').cast(IntegerType()))

card\_df=card\_df.withColumn('card\_id', col('card\_id').cast(IntegerType())).withColumn('disp\_id', col('disp\_id').cast(IntegerType()))

clients\_df=clients\_df.withColumn('client\_id', col('client\_id').cast(IntegerType())).withColumn('district\_id', col('district\_id').cast(IntegerType()))

disp\_df=disp\_df.withColumn('client\_id', col('client\_id').cast(IntegerType())).withColumn('disp\_id', col('disp\_id').cast(IntegerType())) \

.withColumn('account\_id', col('account\_id').cast(IntegerType()))

loan\_df=loan\_df.withColumn('loan\_id', col('loan\_id').cast(IntegerType())).withColumn('account\_id', col('account\_id').cast(IntegerType()))

trans\_df=trans\_df.withColumn('trans\_id', col('trans\_id').cast(IntegerType())).withColumn('account\_id', col('account\_id').cast(IntegerType()))

district\_df = district\_df.withColumn('A1', col('A1').cast(IntegerType()))











