ScholarNest

Spark Azure Databricks

Databricks Spark Certification and beyond

Instructor: Prashant Kumar Pandey

Azure
Databricks

Absolute Beginner to Specialization in Apache Spark and Azure Databricks

Assignments

Assignment - 1

Create a Data frame using the following data file.

/databricks-datasets/airlines/part-00000

Take a stratified sample that meets the following requirement.

1. Create a sample data frame of a total of 10K records
2. Sample must be taken for the DL and PS UniqueCarrier codes
3. Each UniqueCarrier should contribute approximately 5K records in the sample

Assignment 2

You are given a data list as shown below.

```
data_list = [(100, "Prashant", "2020-06-15", 9238614990, "12000", 18.5),
             (101, "David", "2018-08-7", 8908617610, "15000", "nil"),
             (102, "Simran", "14-05-2019", None, "3000000000", 21)]
```

The schema for the above data list is given as below.

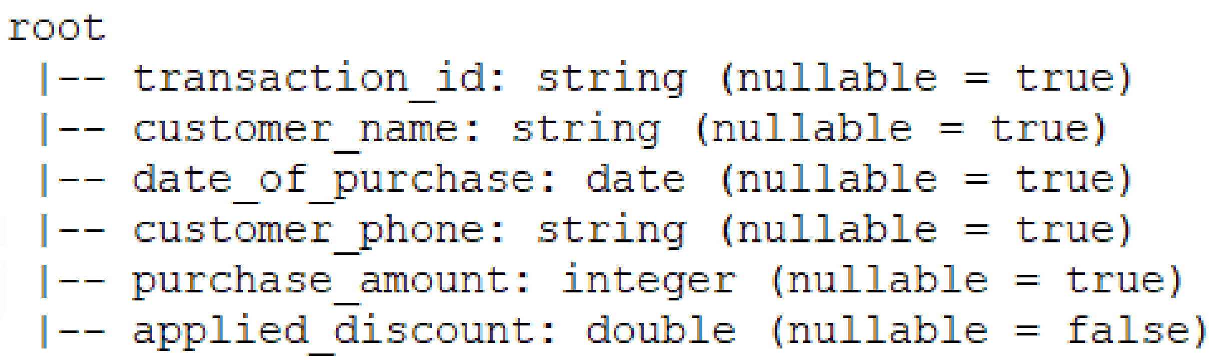
```
schema = 'id int, name string, dop string, phone long, amount string, discount string’
```

Do the following.

1. Create a data frame using the above data list.
2. Convert id from integer to string and rename it as transaction_id
3. Rename the name column to customer_name
4. Convert the dop to date format and rename the column to date_of_purchase
5. You have dates in YYYY-MM-DD and DD-MM-YYYY. Transform them into standard YYYY-MM-DD format
6. Rename the phone column to customer_phone and convert it to string
7. Convert the amount to an Integer value and filter out nulls and outlier values. Rename the column to purchase_amount
8. Convert discount to double, convert nil and null values to zero. Rename the column to applied_discount

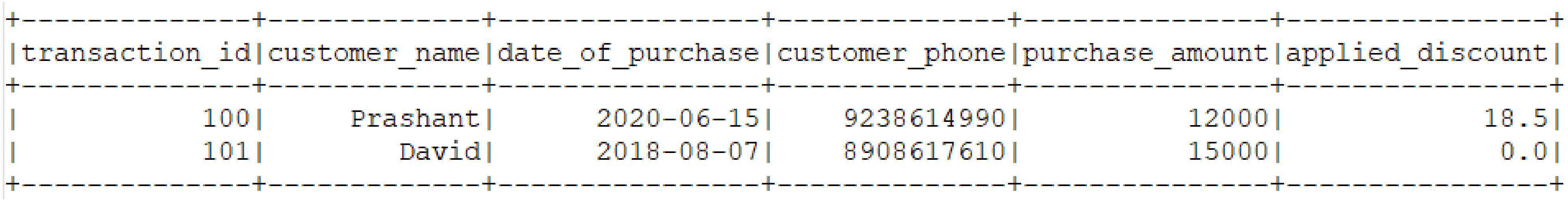
Assignment 2

1. Print the Schema of the new data frame and it should match the below schema.



```
root
 |-- transaction_id: string (nullable = true)
 |-- customer_name: string (nullable = true)
 |-- date_of_purchase: date (nullable = true)
 |-- customer_phone: string (nullable = true)
 |-- purchase_amount: integer (nullable = true)
 |-- applied_discount: double (nullable = false)
```

2. Show the final data frame and the result should match the below output.



```
+--------------+-------------+----------------+--------------+---------------+----------------+
|transaction_id|customer_name|date_of_purchase|customer_phone|purchase_amount|applied_discount|
+--------------+-------------+----------------+--------------+---------------+----------------+
|           100|     Prashant|      2020-06-15|    9238614990|          12000|            18.5|
|           101|        David|      2018-08-07|    8908617610|          15000|             0.0|
+--------------+-------------+----------------+--------------+---------------+----------------+
```

Thank You

ScholarNest Technologies Pvt Ltd.

www.scholarnest.com