## Technology Stack

Kafka, Spark, Scala, SparkSQL, Spark DataFrames, Parquet

## Fields (Schema)

|  |  |
| --- | --- |
| **Field name** | **Description** |
| order\_id | Order Id |
| customer\_id | Customer Id |
| customer\_name | Customer Name |
| product\_id | Product Id |
| product\_name | Product Name |
| product\_category | Product Category |
| payment\_type | Payment Type (card, Internet Banking, UPI, Wallet) |
| qty | Quantity ordered |
| price | Price of the product |
| datetime | Date and time when order was placed |
| country | Customer Country |
| city | Customer City |
| ecommerce\_website\_name | Site from where order was placed |
| payment\_txn\_id | Payment Transaction Confirmation Id |
| payment\_txn\_success | Payment Success or Failure (Y=Success. N=Failed) |
| failure\_reason | Reason for payment failure |

## Sample Data (CSV)

1,101,John Smith,201,Pen,Stationery,Card,24,10,2021-01-10 10:12,India,Mumbai,www.amazon.com,36766,Y,

2,102,Mary Jane,202,Pencil,Stationery,Internet Banking,36,5,2021-10-31 13:45,USA,Boston,www.flipkart.com,37167,Y,

3,103,Joe Smith,203,Some mobile,Electronics,UPI,1,4999,2021-04-23 11:32,UK,Oxford,www.tatacliq.com,90383,Y,

4,104,Neo,204,Some laptop,Electronics,Wallet,1,59999,2021-06-13 15:20,India,Indore,www.amazon.in,12224,N,Invalid CVV.

5,105,Trinity,205,Some book,Books,Card,1,259,2021-08-26 19:54,India,Bengaluru,www.ebay.in,99958,Y,

**Tasks**:

1. Create a producer program in Python that will ingest data to a Kafka Topic.
   1. Data will have to be generated in the program.
   2. Ingest the data every 2 seconds into the Kafka Topic.
2. Display the data from the input Kafka Topic in a console consumer (CLI).
3. Create a consumer program in Python that will read the data stream from the input Kafka Topic and will process the data further.
   1. Read the data into a DataFrame object.
   2. Print the schema of the input data stream
   3. Apply the [above-mentioned schema](#_Fields_(Schema)) to the dataframe and print the schema.
   4. Categorize the data as follows:
      1. Payment types: Card, Internet Banking, UPI, Wallet, Google PAY, PAYTM etc.
      2. Success and Failed payment transactions.
   5. Create separate topics for each of the following and send respective data rows to them:
      1. Card.
      2. Internet Banking.
      3. UPI.
      4. Wallet.
      5. Successful Transactions.
      6. Failed Transactions.
   6. From the consumer program:
      1. Determine and display on the console the number of orders and total amount for each city and payment type.
      2. Also write the same information to a Parquet file (*data should be appended to this file*).