Practice Problem

Welcome to this session on Practice Problems on Producer and Consumer. In this session, you will be given some practice problems related to Producer and Consumer APIs. You will first need to read the data from a CSV file and then write the data into a Kafka topic. Once that is done then you will write codes to read data from these topics and then filter the data based on the given conditions and then write it to some other Kafka topic.

Before you proceed ahead with the practice problems, please download the dataset using this link on your local system. Once, you have downloaded the file, you need to extract it. Name the CSV file present in this folder as **"1500000SalesRecords.csv"**. Transfer this file to your EC2 instance.

The details to transfer the file from your local system to EC2 instance is attached below.

File Transfer - Windows
Download
File Transfer - Linux/Mac
Download

The data contains details of sales orders for regions of Europe. Data schema for the data set is as follows:
Region, Country, Item Type, Sales Channel, Order Priority, Order Date, Order ID, Ship Date, Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, Total Profit

Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
Sub-Saharan Afr	South Africa	Fruits	Offline	M	7/27/2012	443368995	7/28/2012	1593	9.33	6.92	14862.69	11023.56	3839.13
Middle East and	Morocco	Clothes	Online	M	9/14/2013	667593514	10/19/2013	4611	109.28	35.84	503890.08	165258.24	338631.84
Australia and Oc	Papua New Guir	Meat	Offline	M	5/15/2015	940995585	6/4/2015	360	421.89	364.69	151880.4	131288.4	20592
Sub-Saharan Afr	Djibouti	Clothes	Offline	H	5/17/2017	880811536	7/2/2017	562	109.28	35.84	61415.36	20142.08	41273.28
Europe	Slovakia	Beverages	Offline	L	10/26/2016	174590194	12/4/2016	3973	47.45	31.79	188518.85	126301.67	62217.18
Asia	Sri Lanka	Fruits	Online	L	11/7/2011	830192887	12/18/2011	1379	9.33	6.92	12866.07	9542.68	3323.39
Sub-Saharan Afr	Seychelles	Beverages	Online	M	1/18/2013	425793445	2/16/2013	597	47.45	31.79	28327.65	18978.63	9349.02
Sub-Saharan Afr	Tanzania	Beverages	Online	L	11/30/2016	659878194	1/16/2017	1476	47.45	31.79	70036.2	46922.04	23114.16
Sub-Saharan Afr	Ghana	Office Supplies	Online	L	3/23/2017	601245963	4/15/2017	896	651.21	524.96	583484.16	470364.16	113120
Sub-Saharan Afr	Tanzania	Cosmetics	Offline	L	5/23/2016	739008080	5/24/2016	7768	437.2	263.33	3396169.6	2045547.44	1350622.16
Asia	Taiwan	Fruits	Offline	M	2/9/2014	732588374	2/23/2014	8034	9.33	6.92	74957.22	55595.28	19361.94
Middle East and	Algeria	Cosmetics	Online	M	2/18/2011	761723172	2/24/2011	9669	437.2	263.33	4227286.8	2546137.77	1681149.03
Asia	Singapore	Snacks	Online	C	1/28/2013	176461303	2/7/2013	7676	152.58	97.44	1171204.08	747949.44	423254.64
Australia and Oc	Papua New Guir	Clothes	Offline	L	6/20/2011	647164094	7/14/2011	9092	109.28	35.84	993573.76	325857.28	667716.48
Asia	Vietnam	Personal Care	Online	M	4/4/2010	314505374	5/6/2010	7984	81.73	56.67	652532.32	452453.28	200079.04
Sub-Saharan Afr	Uganda	Personal Care	Online	M	6/19/2014	539471471	7/21/2014	451	81.73	56.67	36860.23	25558.17	11302.06
Sub-Saharan Afr	Zimbabwe	Office											

Sample Data Set

Let's now look at the problems in detail.

Problem 1

Write a Kafka producer that will read "**1500000SalesRecords.csv**" file line by line and convert each line to JSON and send the data to a Kafka topic. The name of the topic should be "**test.topic.raw**".

After you have successfully written the Producer Application, you can make use of the Kafka Consumer Console to check whether the data was written to the topic or not. The command to start the consumer console for the above topic is as follows.

bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test.topic.raw --from-beginning

Note: Make sure to run the above command from the Kafka directory.

When you run this command it should get the data in the format as shown below:

{"Region": "Europe", "Country": "Russia", "Item Type": "Cereal", "Sales Channel": "Offline", "Order Priority": "L", "Order Date": "12/21/2012", "Order ID": "874472081", "Ship Date": "12/30/2012", "Units Sold": "4576", "Unit Price": "205.70", "Unit Cost": "117.11", "Total Revenue": "941283.20", "Total Cost": "535895.36", "Total Profit\n": "405387.84\n"}
{"Region": "Central America and the Caribbean", "Country": "Dominica", "Item Type": "Cosmetics", "Sales Channel": "Online", "Order Priority": "M", "Order Date": "3/28/2017", "Order ID": "417365729", "Ship Date": "4/30/2017", "Units Sold": "6649", "Unit Price": "437.20", "Unit Cost": "263.33", "Total Revenue": "2906942.80", "Total Cost": "1750881.17", "Total Profit\n": "1156061.63\n"}
{"Region": "North America", "Country": "United States of America", "Item Type": "Fruits", "Sales Channel": "Offline", "Order Priority": "M", "Order Date": "11/19/2013", "Order ID": "666419351", "Ship Date": "12/13/2013", "Units Sold": "7150", "Unit Price": "9.33", "Unit Cost": "6.92", "Total Revenue": "66709.50", "Total Cost": "49478.00", "Total Profit\n": "17231.50\n"}
{"Region": "Europe", "Country": "Bosnia and Herzegovina", "Item Type": "Snacks", "Sales Channel": "Offline", "Order Priority": "H", "Order Date": "2/13/2011", "Order ID": "943556773", "Ship Date": "3/6/2011", "Units Sold": "8272", "Unit Price": "152.58", "Unit Cost": "97.44", "Total Revenue": "1262141.76", "Total Cost": "806023.68", "Total Profit\n": "456118.08\n"}
{"Region": "Sub-Saharan Africa", "Country": "Comoros", "Item Type": "Clothes", "Sales Channel": "Offline", "Order Priority": "M", "Order Date": "8/27/2010", "Order ID": "191014325", "Ship Date": "9/21/2010", "Units Sold": "4238", "Unit Price": "109.28", "Unit Cost": "35.84", "Total Revenue": "463128.64", "Total Cost": "151889.92", "Total Profit\n": "311238.72\n"}
{"Region": "Australia and Oceania", "Country": "Tonga", "Item Type": "Fruits", "Sales Channel": "Online", "Order Priority": "M", "Order Date": "6/20/2011", "Order ID": "121340835", "Ship Date": "7/15/2011", "Units Sold": "6304", "Unit Price": "9.33", "Unit Cost": "6.92", "Total Revenue": "58816.32", "Total Cost": "43623.68", "Total Profit\n": "15192.64\n"}
{"Region": "Middle East and North Africa", "Country": "United Arab Emirates", "Item Type": "Fruits", "Sales Channel": "Online", "Order Priority": "M", "Order Date": "8/10/2013", "Order ID": "267397630", "Ship Date": "9/3/2013", "Units Sold": "7756", "Unit Price": "9.33", "Unit Cost": "6.92", "Total Revenue": "72363.48", "Total Cost": "53671.52", "Total Profit\n": "18691.96\n"}
{"Region": "Asia", "Country": "North Korea", "Item Type": "Snacks", "Sales Channel": "Offline", "Order Priority": "H", "Order Date": "5/27/2014", "Order ID": "232358539", "Ship Date": "6/9/2014", "Units Sold": "8910", "Unit Price": "152.58", "Unit Cost": "97.44", "Total Revenue": "1359487.80", "Total Cost": "868190.40", "Total Profit\n": "491297.40\n"}
{"Region": "Sub-Saharan Africa", "Country": "Gabon", "Item Type": "Personal Care", "Sales Channel": "Online", "Order Priority": "C", "Order Date": "3/29/2012", "Order ID": "139709889", "Ship Date": "5/16/2012", "Units Sold": "2116", "Unit Price": "81.73", "Unit Cost": "56.67", "Total Revenue": "172940.68", "Total Cost": "119913.72", "Total Profit\n": "53026.96\n"}
{"Region": "Middle East and North Africa", "Country": "Egypt", "Item Type": "Office Supplies", "Sales Channel": "Offline", "Order Priority": "C", "Order Date": "5/16/2010", "Order ID": "183250510", "Ship Date": "6/4/2010", "Units Sold": "6753", "Unit Price": "651.21", "Unit Cost": "524.96", "Total Revenue": "4397621.13", "Total Cost": "3545054.88", "Total Profit\n": "852566.25\n"}
{"Region": "Australia and Oceania", "Country": "Tonga", "Item Type": "Baby Food", "Sales Channel": "Online", "Order Priority": "H", "Order Date": "2/4/2013", "Order ID": "177091729", "Ship Date": "3/11/2013", "Units Sold": "4233", "Unit Price": "255.28", "Unit Cost": "159.42", "Total Revenue": "1080600.24", "Total Cost": "674824.86", "Total Profit\n": "405775.38\n"}
{"Region": "Europe", "Country": "Andorra", "Item Type": "Baby Food", "Sales Channel": "Offline", "Order Priority": "C", "Order Date": "6/7/2014", "Order ID": "905657374", "Ship Date": "7/14/2014", "Units Sold": "2998", "Unit Price": "255.28", "Unit Cost": "159.42", "Total Revenue": "765329.44", "Total Cost": "477941.16", "Total Profit\n": "287388.28\n"}
{"Region": "Asia", "Country": "Brunei", "Item Type": "Snacks", "Sales Channel": "Offline", "Order Priority": "L", "Order Date": "1/10/2017", "Order ID": "973683774", "Ship Date": "2/14/2017", "Units Sold": "3521", "Unit Price": "152.58", "Unit Cost": "97.44", "Total Revenue": "537234.18", "Total Cost": "343086.24", "Total Profit\n": "194147.94\n"}
{"Region": "North America", "Country": "United States of America", "Item Type": "Baby Food", "Sales Channel": "Online", "Order Priority": "H", "Order Date": "3/29/2013", "Order ID": "856219756", "Ship Date": "4/22/2013", "Units Sold": "3308", "Unit Price": "255.28", "Unit Cost": "159.42", "Total Revenue": "844466.24", "Total Cost": "527361.36", "Total Profit\n": "317104.88\n"}
}

Data Format

Problem 2

Write a Kafka consumer that will read data from the topic named "**test.topic.raw**". It should read the data from the beginning and do the below filtering of data:

- Select only those data whose "Total Cost" value is greater than 304017.56
- After selection, it should send the filtered data to another topic named "**test.topic.threshold_costs**"

After successfully executing this code, use the following command to see the messages present in the topic named "test.topic.threshold_costs".

bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic test.topic.threshold_costs --from-beginning

Note: Make sure to run the above command from the Kafka directory.

When you run this command it should get the data in the format as shown below:

["Region": "Europe", "Country": "Serbia", "Item Type": "Cosmetics", "Sales Channel": "Online", "Order Priority": "H", "Order Date": "10/10/2012", "Order ID": "984903562", "Ship Date": "11/1
3/2012", "Units Sold": "3705", "Unit Price": "437.20", "Unit Cost": "263.33", "Total Revenue": "1619826.00", "Total Cost": "975637.65", "Total Profit\n": "644188.35\n"]
["Region": "Central America and the Caribbean", "Country": "Panama", "Item Type": "Household", "Sales Channel": "Offline", "Order Priority": "M", "Order Date": "10/20/2016", "Order ID": "33
1746184", "Ship Date": "11/27/2016", "Units Sold": "9923", "Unit Price": "668.27", "Unit Cost": "502.54", "Total Revenue": "6631243.21", "Total Cost": "4986704.42", "Total Profit\n": "16445
38.79\n"]
["Region": "Asia", "Country": "Mongolia", "Item Type": "Cereal", "Sales Channel": "Online", "Order Priority": "M", "Order Date": "5/15/2015", "Order ID": "425405490", "Ship Date": "6/26/201
5", "Units Sold": "5793", "Unit Price": "205.70", "Unit Cost": "117.11", "Total Revenue": "1191620.10", "Total Cost": "678418.23", "Total Profit\n": "513201.87\n"]
["Region": "Sub-Saharan Africa", "Country": "Rwanda", "Item Type": "Vegetables", "Sales Channel": "Offline", "Order Priority": "L", "Order Date": "3/4/2013", "Order ID": "294066512", "Ship
Date": "3/20/2013", "Units Sold": "9671", "Unit Price": "154.06", "Unit Cost": "90.93", "Total Revenue": "1489914.26", "Total Cost": "879384.03", "Total Profit\n": "610530.23\n"]
["Region": "Sub-Saharan Africa", "Country": "Nigeria", "Item Type": "Clothes", "Sales Channel": "Offline", "Order Priority": "M", "Order Date": "6/1/2017", "Order ID": "364317643", "Ship Da
te": "6/16/2017", "Units Sold": "8541", "Unit Price": "109.28", "Unit Cost": "35.84", "Total Revenue": "933360.48", "Total Cost": "306109.44", "Total Profit\n": "627251.04\n"]
["Region": "Middle East and North Africa", "Country": "Afghanistan", "Item Type": "Cereal", "Sales Channel": "Offline", "Order Priority": "H", "Order Date": "2/8/2014", "Order ID": "6683166
38", "Ship Date": "3/16/2014", "Units Sold": "8600", "Unit Price": "205.70", "Unit Cost": "117.11", "Total Revenue": "1769020.00", "Total Cost": "1007146.00", "Total Profit\n": "761874.00\n
"]
["Region": "Australia and Oceania", "Country": "Fiji", "Item Type": "Vegetables", "Sales Channel": "Offline", "Order Priority": "H", "Order Date": "11/18/2014", "Order ID": "204549825", "Sh
ip Date": "12/7/2014", "Units Sold": "9694", "Unit Price": "154.06", "Unit Cost": "90.93", "Total Revenue": "1493457.64", "Total Cost": "881475.42", "Total Profit\n": "611982.22\n"]
["Region": "Sub-Saharan Africa", "Country": "Equatorial Guinea", "Item Type": "Household", "Sales Channel": "Offline", "Order Priority": "H", "Order Date": "7/8/2013", "Order ID": "74181603
1", "Ship Date": "8/23/2013", "Units Sold": "8703", "Unit Price": "668.27", "Unit Cost": "502.54", "Total Revenue": "5815953.81", "Total Cost": "4373605.62", "Total Profit\n": "1442348.19\n
"]
["Region": "Asia", "Country": "South Korea", "Item Type": "Cosmetics", "Sales Channel": "Online", "Order Priority": "M", "Order Date": "12/6/2010", "Order ID": "552488553", "Ship Date": "1/
13/2011", "Units Sold": "2495", "Unit Price": "437.20", "Unit Cost": "263.33", "Total Revenue": "1090814.00", "Total Cost": "657008.35", "Total Profit\n": "433805.65\n"]
["Region": "Europe", "Country": "Monaco", "Item Type": "Office Supplies", "Sales Channel": "Offline", "Order Priority": "L", "Order Date": "2/16/2013", "Order ID": "673476350", "Ship Date":
"3/5/2013", "Units Sold": "5473", "Unit Price": "651.21", "Unit Cost": "524.96", "Total Revenue": "3564072.33", "Total Cost": "2873106.08", "Total Profit\n": "690966.25\n"]
["Region": "Central America and the Caribbean", "Country": "Cuba", "Item Type": "Baby Food", "Sales Channel": "Offline", "Order Priority": "M", "Order Date": "3/23/2011", "Order ID": "63019
1934", "Ship Date": "5/7/2011", "Units Sold": "6652", "Unit Price": "255.28", "Unit Cost": "159.42", "Total Revenue": "1698122.56", "Total Cost": "1060461.84", "Total Profit\n": "637660.72\
n"]

Data Format

Problem 3

Write a Kafka consumer that will read data from multiple topics:

- The name of the topics is "**test.topic.raw**" and "**test.topic.threshold_costs**"
- They should also read data from the **beginning** of the topic
- Filter the data based on the country and send data to different topic depending on country value "**test.topic.{country}**". For each country, a new topic needs to be created and data needs to be sent to the corresponding topics. So, if there is some data for the country Cuba, then a topic named **test.topic.cuba** should be created and the data should be sent to this topic. Similarly for other countries. You need not create separate topics for each country using the command line. Think of some logic to implement it using python code only.
- When you list all the topics, you will see that the topics are created for each country

The command to list all the topics is as follows.

bin/kafka-topics.sh --bootstrap-server localhost:9092 --list

When you run this command you will see the lists of different topics being created. There will be a topic for each country.

```
test.topic.Cambodia
test.topic.Canada
test.topic.CapeVerde
test.topic.CentralAfricanRepublic
test.topic.China
test.topic.CzechRepublic
test.topic.DemocraticRepublicoftheCongo
test.topic.Denmark
test.topic.Dominica
test.topic.EastTimor
test.topic.ElSalvador
test.topic.EquatorialGuinea
test.topic.Estonia
test.topic.Finland
test.topic.France
test.topic.Ghana
test.topic.Greece
test.topic.Greenland
test.topic.Haiti
test.topic.Iceland
test.topic.India
test.topic.Indonesia
test.topic.Italy
test.topic.Japan
test.topic.Kazakhstan
test.topic.Kenya
test.topic.Kuwait
test.topic.Latvia
test.topic.Lebanon
test.topic.Madagascar
test.topic.Malawi
test.topic.Mali
test.topic.MarshallIslands
test.topic.Mongolia
test.topic.Morocco
test.topic.Myanmar
test.topic.Nepal
test.topic.Netherlands
test.topic.Nicaragua
test.topic.Oman
test.topic.Pakistan
test.topic.PapuaNewGuinea
test.topic.Portugal
test.topic.Romania
```

Topic List

Problem 4

Write down 2 Kafka consumers, both reading data from the same topic
"test.topic.raw". Both consumers should receive the same messages and do as
described ahead.

Consumer1 should do the following:
- Get the "**Sale Channel**" from JSON data
- Send data to a Kafka topic depending on the channel of the sale. If for a sale the channel is offline then that data should be sent to a topic named **test.topic.offline**. So, for each unique sale channel, a separate topic has to be created
- List the topics, you should see all the topics created for each Sale Channel

As you can see in the image below topics named **test.topic.offline** and **test.topic.online** has been created.

```
test.topic.NorthKorea
test.topic.Offline
test.topic.Oman
test.topic.Online
test.topic.Pakistan
```

Topics

Consumer2 should do the following:
- Filter the data based on Unit Cost
- Send data to Kafka topic named "**topic.costly_order**" if 'Unit Cost' is more than "**10000**"
- Start consumer console to verify data reaching to the given topic.