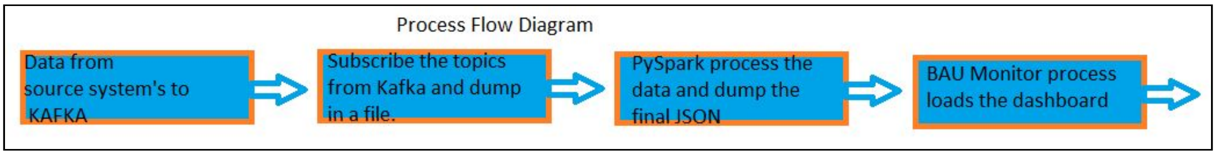
Data Streaming Application in kafka and Spark Processing

This is a standalone application for stream data processing from source, process the data in pyspark and load the data as JSON in the local directory..

The data pipeline is given below,



Process Flow Diagram

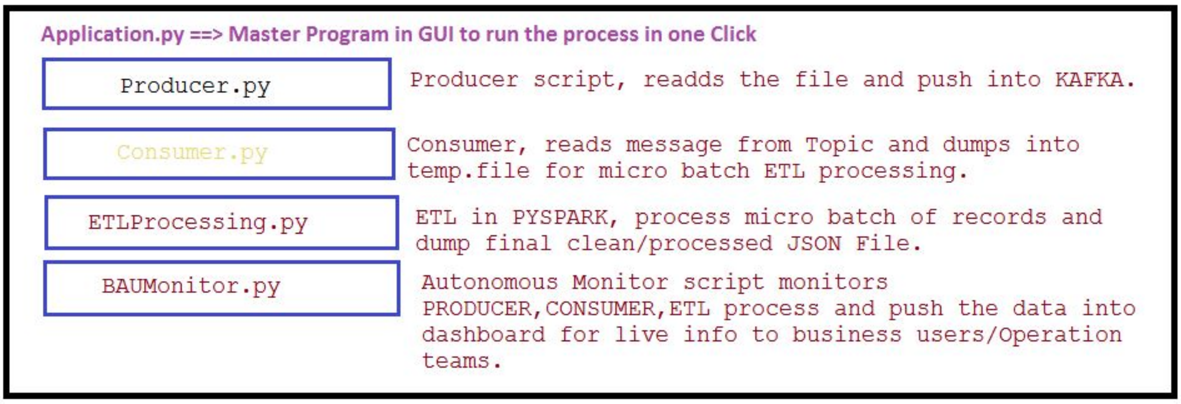
Data from source system's to KAFKA

Subscribe the topics from Kafka and dump in a file.

PySpark process the data and dump the final JSON

BAU Monitor process loads the dashboard

Script Flow Diagram:

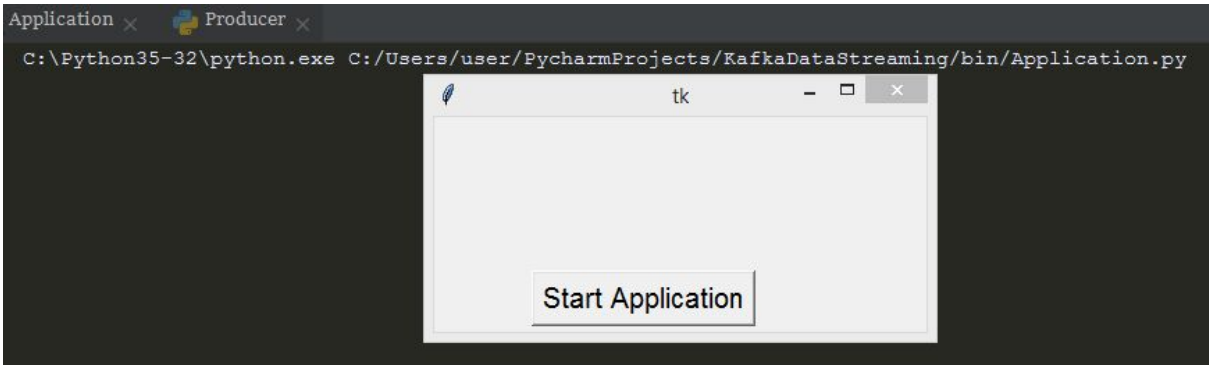


Application.py ==> Master Program in GUI to run the process in one Click

Producer.py	Producer script, readds the file and push into KAFKA.
Consumer.py	Consumer, reads message from Topic and dumps into temp.file for micro batch ETL processing.
ETLProcessing.py	ETL in PYSPARK, process micro batch of records and dump final clean/processed JSON File.
BAUMonitor.py	Autonomous Monitor script monitors PRODUCER,CONSUMER,ETL process and push the data into dashboard for live info to business users/Operation teams.

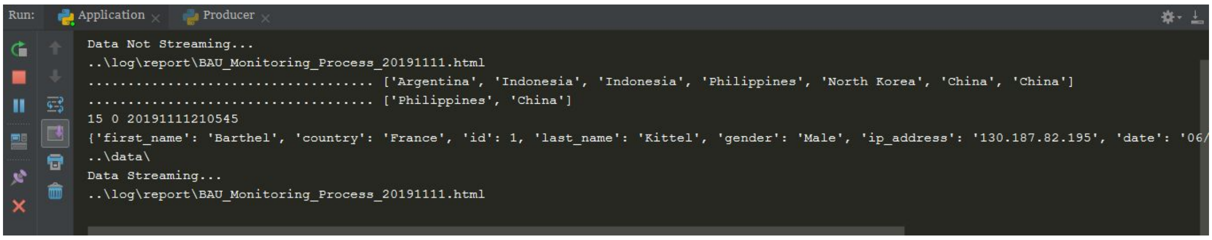
Execution Steps:

User just need to execute the Applciation.py a GUI opens and should Start application as below,



```
C:\Python35-32\python.exe C:/Users/user/PycharmProjects/KafkaDataStreaming/bin/Application.py
```

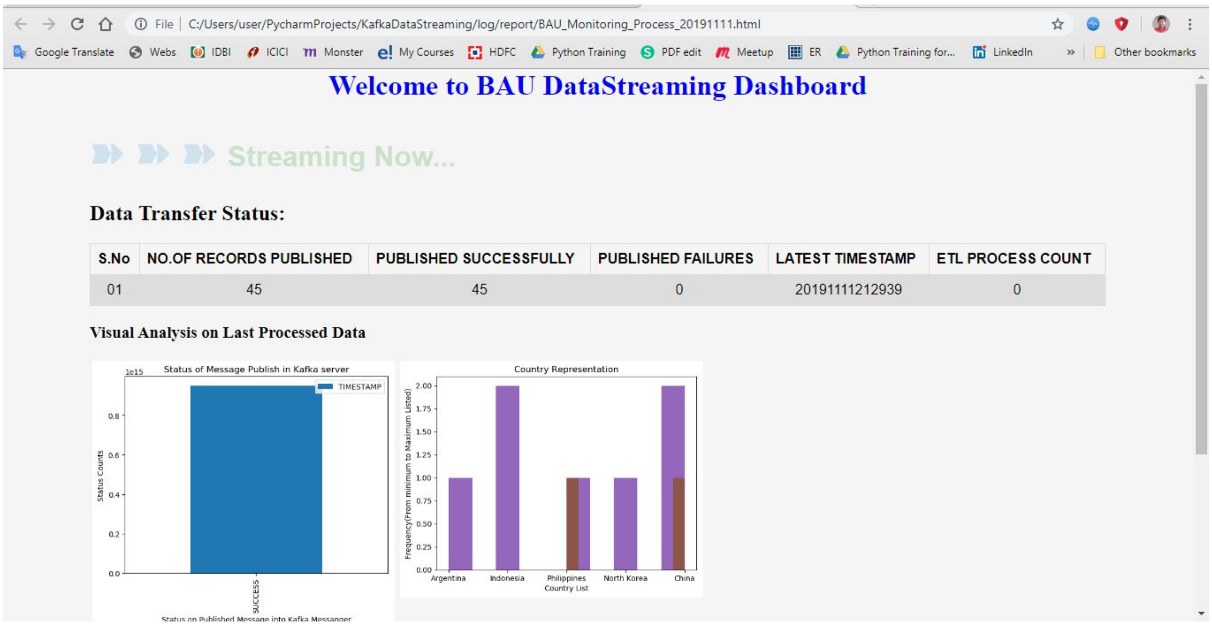
Upon clicking the Start application,process starts as,



```
Data Not Streaming...
..\log\report\BAU_Monitoring_Process_20191111.html
.................................... ['Argentina', 'Indonesia', 'Indonesia', 'Philippines', 'North Korea', 'China', 'China']
.................................... ['Philippines', 'China']
15 0 20191111210545
{'first_name': 'Barthel', 'country': 'France', 'id': 1, 'last_name': 'Kittel', 'gender': 'Male', 'ip_address': '130.187.82.195', 'date': '06/
..\data\
Data Streaming...
..\log\report\BAU_Monitoring_Process_20191111.html
```

Dashboard View:

Since this is just a temporary project parameters needs to be tuned fined for excellent performance and better GUI/UI/UX performance. For now,the Dashboard(HTML) should be manually refreshed over browser due to the limitation of my personal laptop configurations, haven't linked to any realtime reporting.



Welcome to BAU DataStreaming Dashboard

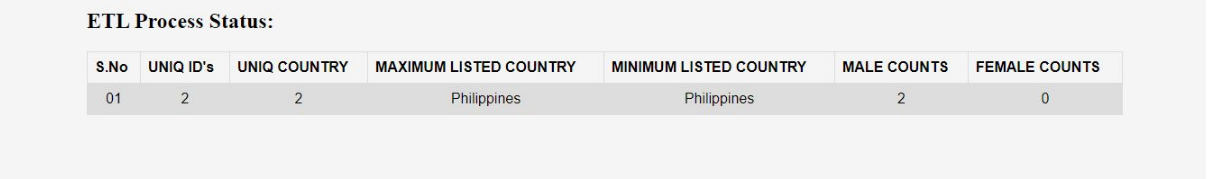
Streaming Now...

Data Transfer Status:

S.No	NO.OF RECORDS PUBLISHED	PUBLISHED SUCCESSFULLY	PUBLISHED FAILURES	LATEST TIMESTAMP	ETL PROCESS COUNT
01	45	45	0	20191111212939	0

Visual Analysis on Last Processed Data

SNAPSHOT 2:



ETL Process Status:

S.No	UNIQ ID's	UNIQ COUNTRY	MAXIMUM LISTED COUNTRY	MINIMUM LISTED COUNTRY	MALE COUNTS	FEMALE COUNTS
01	2	2	Philippines	Philippines	2	0

The code is designed so robust and versatile in extending to other applications/further integrations with new softwares/framework.Almost the code is kind of plug-play!

NOTE:

ETL process has been segregated so that complex logic can be handled separately and actual data streaming cycle remains un disturbed incase of time dependent streaming from Multiple data sources.

Data process happens in multiple micro batches to process the data efficiently using the best of the available resources/timekeeping in the fact that downstream receives good amount of data for their incremental data process.