Servify Assignment

Task: Ingest the real-time data to Kafka and process the following metrics:

1. The average number of plans sold per week.
2. The brand (BrandID) with the highest number of plans bought by customers.
3. The percentage of service requests raised under a plan of the total number of requests raised.

Approach:

- The ingestion or first part -
    - **How I am doing it?**
        - **Ingestion**: I am using python for this and pymysql library(for connection) and making a connection string and taking the current snapshot of it(it's not a real time but still) and directly using the producer API of kafka, I am ingesting it as a json. I am also making topic for every table given in the task(consumer-topic, consumer_product-topic, consumer_servicerequest-topic, sold_plan-topic)

```
producer = KafkaProducer(bootstrap_servers = 'localhost:9092')
for row in consumer_results:
    producer.send('consumer-topic', json.dumps(row, default=str).encode())

for row in consumer_product_results:
    producer.send('consumer_product-topic', json.dumps(row, default=str).encode())

for row in consumer_servicerequest_results:
    producer.send('consumer_servicerequest-topic', json.dumps(row, default=str).encode())

for row in sold_plan_results:
    producer.send('sold_plan-topic', json.dumps(row, default=str).encode())
```

    - **Other Way:**
        - There is another method I thought(more real time than the snapshot or the first one) using the connector API of Kafka. As I have given the remote MySQL server I can directly pass the URL and it will directly fetch the data from the remote server and ingest it to Kafka.
- Pre Processing step:
    - **How I am doing it?**
        - **Cleaning**: I am not doing the cleaning step(I assumed the data I am getting is fine). But as I am fetching the all the data in a list and then row by row ingesting it into kafka I can directly check the row for null values and can drop the row there and doesn't ingest it into kafka.

- Stream for metrics:
    - **How I am doing it?**
        - This task consist of two steps:
            - **Streaming using either Kafka Streams or KSQL(Not done with it yet)**: This is bit tricky as streams API wrapper for python is not there(there are some individual clients and there is a issue in confluent-kafka-python repo which is still in progress (https://bit.ly/2AF9xUB). So I have to go for java. I also tried my hands with KSQL but it is still pretty new and the documentation is still not mature so I thought give a try to stream API instead of it.
            - **Metrics(Not done with it yet)**: New to me and I am trying best to find out solution

Learning:

- **Streaming**: I never used streams API before this assignment. I learnt about it. As well as I got to know about the connect API(as I thought of making this more real time)
- **Metrics**
- **Back to Java**: These days I am coding more in Python instead of Java and for this assignment I have to go back to java which is refreshing for me.

Resources/References:

- Series of Yelp real time data pipeline(https://bit.ly/2zGXJBm)
- Apache Kafka Documentation
- Kafka Streams Webinar by a channel Big-Engineering(https://bit.ly/2AH157q)
- A talk by Jay Kreps in Reactive summit 2016. Good for me to get started and get a bit of overview on Streams. (https://bit.ly/2hw8G4e)

For Deployment(Not done):

- **JAR File**: In Java, I can export files into JAR so that they deploy directly to server without any other dependencies.
- **Containerization**: Using Docker, I can containerize the whole assignment and this would be better for deployment directly to the server or AWS.