# **Mistplay Data Engineer Take Home Challenge**

**Main Design**



**1)** On Local machine Mac with below tools and techs installed:-

Spark

Kafka

Python

Anaconda Spyder

Airflow

Docker

**2)** Clone the fork repo with git and create a feature branch

Then the design is divided in to modules independent of each other , starting with Kafka layer

**Kafka Layer**

**3)** Since i want to stream the data from the files

Start Zookeeper

Start kafka Server

Created 5 Topics (Customer, Customer\_Extended, Sales, Refund, Product)   
(In the Producer.py file)

Kafka is all set and running.

**4)** Next I want to Publish my 5 text file data to kafka using Kafka-Producer()

5 files published to Customer, Customer\_Extended, Sales, Refund, Product Topics

**5)** To execute the step-4, i wrote python script **Producer.py** which will read the text files and pushes it to Kafka-Producer using python Library (KafkaProducer of kafka).

**6)** After Step-5 the data’s are streaming via respective topics.

**Spark streaming Tranformation Layer**

**7)** Now I want to capture these data , do some transformations and load it on staging layer.

**8)** To execute Step-7 I used spark-Structured-Streaming (Spark-Consumer) to capture it , for this i used Pyspark(python Library) in **Consumer.py** script

**10)** Now my records are ready in staging tables , i want to write some sql queries and want to see some insights using charts.

**Analytics Layer**

**11)** To execute Step-10 i used pandas to get insights and answer the questions

**12)** Total 3 Python Jobs.

→ One for Kafka-Producer (JOB1)

→ One for spark-Streaming consumer (JOB2)

→ one for Sql queries and insights (JOB3)

**Dockerizing the whole solution**

Below are the Docker(**docker-compose.yaml** file) Containers i am using it:

**Postgresql for Airflow db , Airflow , Zookeeper for Kafka server, Kafka Server, Spark**

We will use Airflow to trigger the 3 jobs mentioned above.

Airflow will trigger JOB1 , JOB2 , JOB3 using airflow dags,

Now to get the better clarity I will make 3 DAGs to trigger 3 jobs

→ DAG1(data\_stream\_DAG.py), This will trigger the job JOB1 of producer

→ DAG2(spark\_consumer\_DAG.py), This will trigger the job JOB2 to consume data and load into tables.

→ DAG3(insight\_DAG.py), This will trigger the job JOB3 to show insights using charts.

**so now we have 4 container in total :-**

Container 1 (Postgresql for Airflow db)

Container 2 (Airflow + KafkaProducer + Spark\_hadoop)

Container 3 (Zookeeper for Kafka server)

Container 4 (Kafka Server)