PG Certificate in Software Engineering for Data Science
KAFKA + SPARK

Problem Overview:

In this Problem Statement we will see how we can use Kafka to understand Stream Processing by receiving tweets from twitter through tweepy API and send it to Kafka server from which the tweets are received by stream processing application Spark which processes the tweets and stores it in the Hive Database.

For demonstration we will use fake tweets by combining random words to show the working of kafka, spark and hive.

Requirements:

The software required for this application are:

1. Kafka
2. Hadoop
3. Hive
4. Spark

The procedure to install all these packages are given in the Requirements file.

Assignment:

1. Fix the tweet_stream.py file with the API keys from your twitter developer account to get tweets from the twitter directly and send it to kafka.
2. Train a sentiment analysis model and use your model in place of a pretrained model to analyze the sentiment of the tweets it receives and store it in the hive.

Submission:

Submit the tweet_stream.py, transformer.py file containing the model that you have trained and used and output of see_contents.py