Assignment 4

You can choose any cloud (AWS, Azure, GCP etc) and spin two instances. You need to install Kafka message queue (zookeeper is also required to be installed) and run Kafka message broker on the two instances. Spark also needs to be installed on these two instances. Installing and running the above packages on the cloud instances is part of the assignment. No separate pdf instructions will be provided for the same. You can download these instructions along with the packages. The virtual box used in the class will have pdf instructions for running but these will be specific to the virtual box only and this can used for learning and doing it on the cloud independently. This is part of the learning in the assignment.
You need to write programs on the above cloud instances to get live tweets from twitter and detect tweets coming from New York and California states and publish them into Kafka broker under the topics NY and CA. You need to write spark streaming program to read the tweets published under the topics and count total number of tweets under each topic in the past 10 minutes. You need to display the total tweet count along with the most frequently occurring words and word combinations (frequent item sets) in the tweets. Use sufficiently low support threshold to detect upto 3 or more frequent item sets. Use of spark mllib frequent item set library is part of the assignment. You need to provide the url where the programs are running for testing the assignment. The web interface should provide options for choosing either NY or CA and based on the option the total tweet count and frequently occurring words and word combinations should get displayed on the screen for the 10 minute interval.