Module 9 – Understanding Apache Kafka and Apache Flume

Case Study I

Case Study: Spam Detection
Domain: Telecom

A telecom software provider is building an application to monitor different telecom components in the production environment. For monitoring purpose, the application relies on log files by parsing the log files and looking for potential warning or exceptions in the logs and reporting them.
The POC we had been working on, for SPAM Detection on the data of telecom operator forum, has been accepted and the stakeholders have asked us to work on the real-time example for predicting SPAM messages.

Tasks

This POC will focus on saved machine learning model for spam prediction with streaming data to do real-time prediction.
You have already developed the model in previous exercise. Now as part of a POC you are required to publish data through Kafka API, before pushing the data to Spark Streaming. In the later part the streaming application will be subscribed to the Kafka topic.

1. Verify the cluster
2. Create a topic in Kafka so that consumers and produces can enqueue/dequeue data respectively from the topic
3. Write the test Kafka consumer and verify that data is sent successfully.
4. Configure a flume agent to use Kafka as the channel and HDFS as the sink
5. Start flume agent and test the output to HDFS
6. Test the complete pipeline

Hint:

Below is the example for Spam Messages:

- HAM: What you doing? How are you?
- SPAM: Sunshine Quiz! Win a super Sony DVD recorder if you can name the capital of Australia? Text MQUIZ to 82277