Module 9 – Understanding Apache Kafka and Apache Flume

Case Study II

Case Study: Telecom Pipeline
Domain: Telecom

There are two large obstacles in collecting metadata from a network as large as India's Big Telecom operator: transporting the sheer volume of data and processing it before the data no longer accurately reflects the state of the network. Fortunately, combining Apache Flume and Apache Kafka using the Kafka pattern provides a means to move data into the Hadoop cluster and readily scale the pipeline to address both transient and persistent spikes in data volume. Company is planning to deploy Flume and Kafka across the network in a geographically distributed architecture that achieves scale and resilience, having been tuned from around 10,000 events per second on initial deployment to 1,000,000 events per second using a three-node Kafka cluster.

Tasks:

You are part of the Telecom Operator's R&D team, which is required to perform a quick POC on the Kafka Flume pipeline to persist data to HDFS and analyze the data through spark streaming.

Dataset:

The data set consists of 100 variables and approx. 100 thousand records containing different variables explaining the attributes of telecom industry and various factors considered important while dealing with customers of telecom industry.