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:

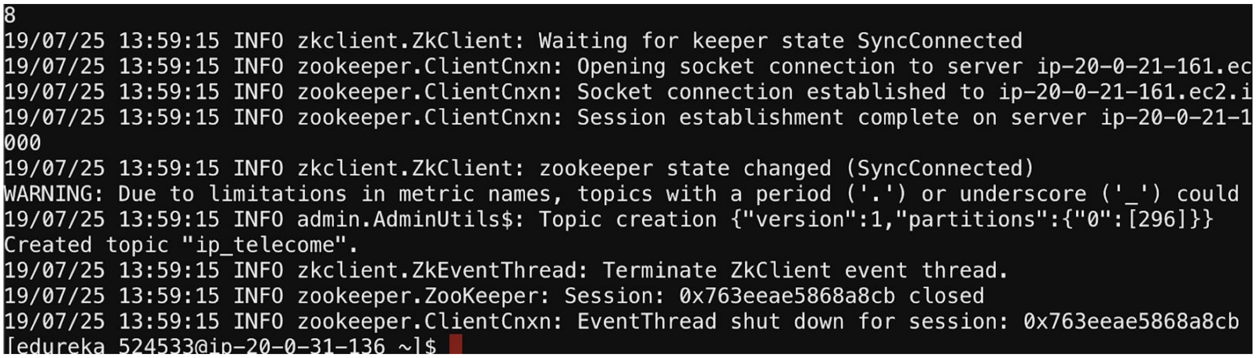
You are part of the Telecom Operator's R&D team, which is required to perform a quick POC on the Kafka Flume pipeline to persist data to HDFS and analyze the data through spark streaming.

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.

Step 1: Create a topic in Kafka so that consumers and produces can enqueue/dequeue data respectively from the topic

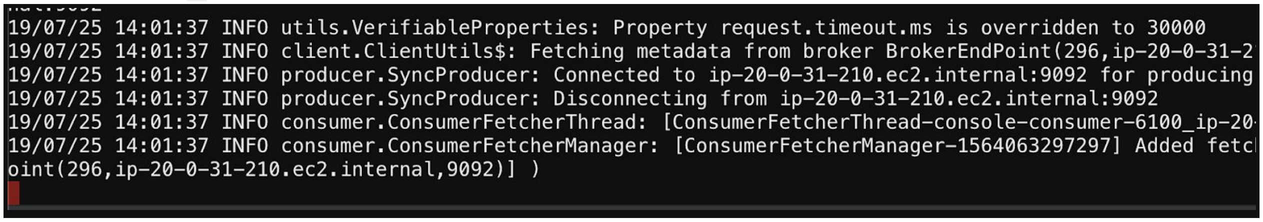
```
kafka-topics --create --zookeeper ip-20-0-21-161.ec2.internal:2181
--replication-factor 1 --partitions 1 --topic ip_telecome
```



```
8
19/07/25 13:59:15 INFO zkclient.ZkClient: Waiting for keeper state SyncConnected
19/07/25 13:59:15 INFO zookeeper.ClientCnxn: Opening socket connection to server ip-20-0-21-161.ec
19/07/25 13:59:15 INFO zookeeper.ClientCnxn: Socket connection established to ip-20-0-21-161.ec2.i
19/07/25 13:59:15 INFO zookeeper.ClientCnxn: Session establishment complete on server ip-20-0-21-1
000
19/07/25 13:59:15 INFO zkclient.ZkClient: zookeeper state changed (SyncConnected)
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could
19/07/25 13:59:15 INFO admin.AdminUtils$: Topic creation {"version":1,"partitions":{"0":[296]}}
Created topic "ip_telecome".
19/07/25 13:59:15 INFO zkclient.ZkEventThread: Terminate ZkClient event thread.
19/07/25 13:59:15 INFO zookeeper.ZooKeeper: Session: 0x763eeae5868a8cb closed
19/07/25 13:59:15 INFO zookeeper.ClientCnxn: EventThread shut down for session: 0x763eeae5868a8cb
[edureka 524533@ip-20-0-31-136 ~]$
```

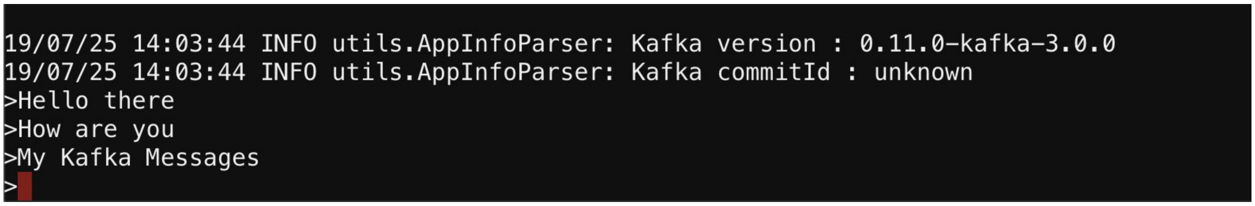
Step 2: Write the test Kafka consumer and verify that data is sent successfully

```
kafka-console-consumer --zookeeper ip-20-0-21-161.ec2.internal:2181
--topic ip_telecome --from-beginning
```



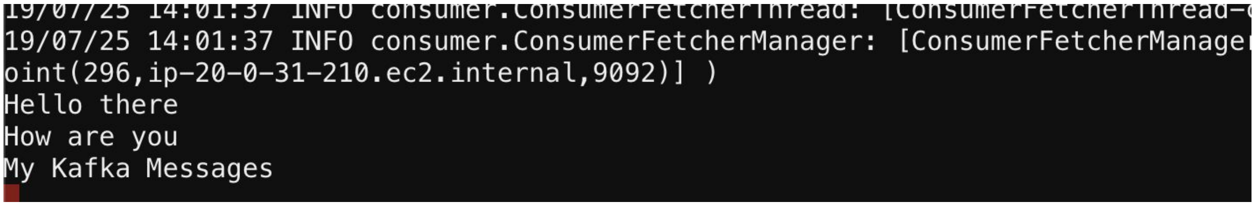
```
19/07/25 14:01:37 INFO utils.VerifiableProperties: Property request.timeout.ms is overridden to 30000
19/07/25 14:01:37 INFO client.ClientUtils$: Fetching metadata from broker BrokerEndPoint(296,ip-20-0-31-2
19/07/25 14:01:37 INFO producer.SyncProducer: Connected to ip-20-0-31-210.ec2.internal:9092 for producing
19/07/25 14:01:37 INFO producer.SyncProducer: Disconnecting from ip-20-0-31-210.ec2.internal:9092
19/07/25 14:01:37 INFO consumer.ConsumerFetcherThread: [ConsumerFetcherThread-console-consumer-6100_ip-20
19/07/25 14:01:37 INFO consumer.ConsumerFetcherManager: [ConsumerFetcherManager-1564063297297] Added fetc
oint(296,ip-20-0-31-210.ec2.internal,9092)] )
```

kafka-console-producer --broker-list ip-20-0-31-4.ec2.internal:9092 --topic ip_telecome



```
19/07/25 14:03:44 INFO utils.AppInfoParser: Kafka version : 0.11.0-kafka-3.0.0
19/07/25 14:03:44 INFO utils.AppInfoParser: Kafka commitId : unknown
>Hello there
>How are you
>My Kafka Messages
>
```

Being received by the Kafka Consumer

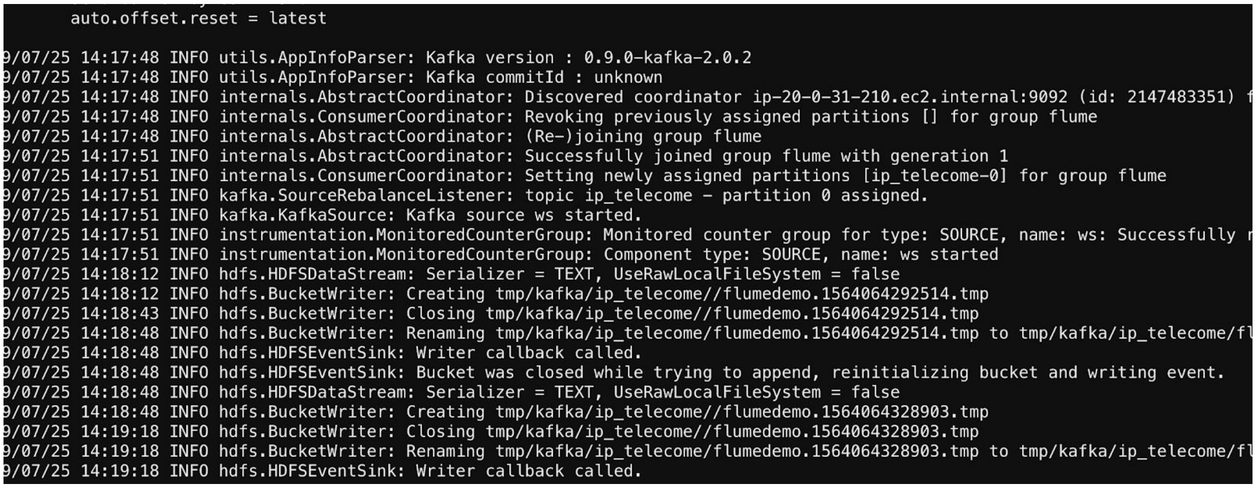


```
19/07/25 14:01:37 INFO consumer.ConsumerFetcherThread: [ConsumerFetcherThread-
19/07/25 14:01:37 INFO consumer.ConsumerFetcherManager: [ConsumerFetcherManage
oint(296,ip-20-0-31-210.ec2.internal,9092)] )
Hello there
How are you
My Kafka Messages
```

Step 3: Configure a flume agent to use Kafka as the channel and HDFS as the sink
Create new file telecom.conf

Step 4: Start flume agent and test the output to HDFS

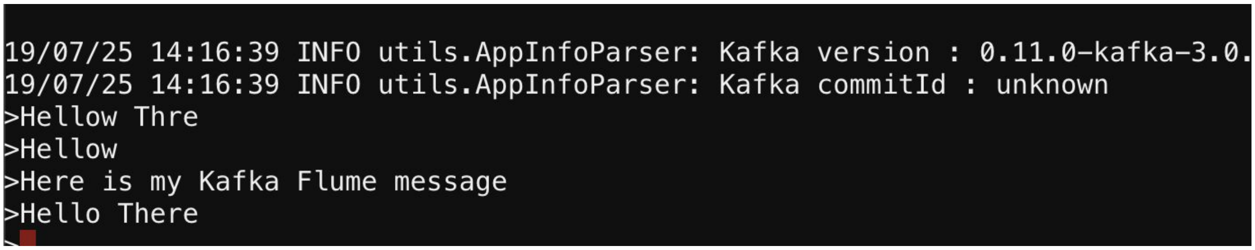
flume-ng agent --conf conf --conf-file telecome.conf --name wh
-Dflume.root.logger=INFO,console



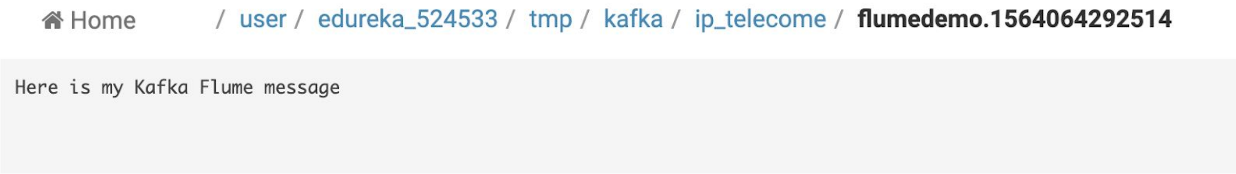
```
      auto.offset.reset = latest

9/07/25 14:17:48 INFO utils.AppInfoParser: Kafka version : 0.9.0-kafka-2.0.2
9/07/25 14:17:48 INFO utils.AppInfoParser: Kafka commitId : unknown
9/07/25 14:17:48 INFO internals.AbstractCoordinator: Discovered coordinator ip-20-0-31-210.ec2.internal:9092 (id: 2147483351) f
9/07/25 14:17:48 INFO internals.ConsumerCoordinator: Revoking previously assigned partitions [] for group flume
9/07/25 14:17:48 INFO internals.AbstractCoordinator: (Re-)joining group flume
9/07/25 14:17:51 INFO internals.AbstractCoordinator: Successfully joined group flume with generation 1
9/07/25 14:17:51 INFO internals.ConsumerCoordinator: Setting newly assigned partitions [ip_telecome-0] for group flume
9/07/25 14:17:51 INFO kafka.SourceRebalanceListener: topic ip_telecome - partition 0 assigned.
9/07/25 14:17:51 INFO kafka.KafkaSource: Kafka source ws started.
9/07/25 14:17:51 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: ws: Successfully r
9/07/25 14:17:51 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: ws started
9/07/25 14:18:12 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
9/07/25 14:18:12 INFO hdfs.BucketWriter: Creating tmp/kafka/ip_telecome//flumedemo.1564064292514.tmp
9/07/25 14:18:43 INFO hdfs.BucketWriter: Closing tmp/kafka/ip_telecome//flumedemo.1564064292514.tmp
9/07/25 14:18:48 INFO hdfs.BucketWriter: Renaming tmp/kafka/ip_telecome/flumedemo.1564064292514.tmp to tmp/kafka/ip_telecome/fl
9/07/25 14:18:48 INFO hdfs.HDFSEventSink: Writer callback called.
9/07/25 14:18:48 INFO hdfs.HDFSEventSink: Bucket was closed while trying to append, reinitializing bucket and writing event.
9/07/25 14:18:48 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
9/07/25 14:18:48 INFO hdfs.BucketWriter: Creating tmp/kafka/ip_telecome//flumedemo.1564064328903.tmp
9/07/25 14:19:18 INFO hdfs.BucketWriter: Closing tmp/kafka/ip_telecome//flumedemo.1564064328903.tmp
9/07/25 14:19:18 INFO hdfs.BucketWriter: Renaming tmp/kafka/ip_telecome/flumedemo.1564064328903.tmp to tmp/kafka/ip_telecome/fl
9/07/25 14:19:18 INFO hdfs.HDFSEventSink: Writer callback called.
```

Step 5: Test the complete pipeline
Check messages in HDFS

hdfs dfs -ls tmp/kafka/ip_telecom



```
19/07/25 14:16:39 INFO utils.AppInfoParser: Kafka version : 0.11.0-kafka-3.0.
19/07/25 14:16:39 INFO utils.AppInfoParser: Kafka commitId : unknown
>Hellow Thre
>Hellow
>Here is my Kafka Flume message
>Hello There
>
```



Home / user / edureka_524533 / tmp / kafka / ip_telecome / **flumedemo.1564064292514**

```
Here is my Kafka Flume message
```