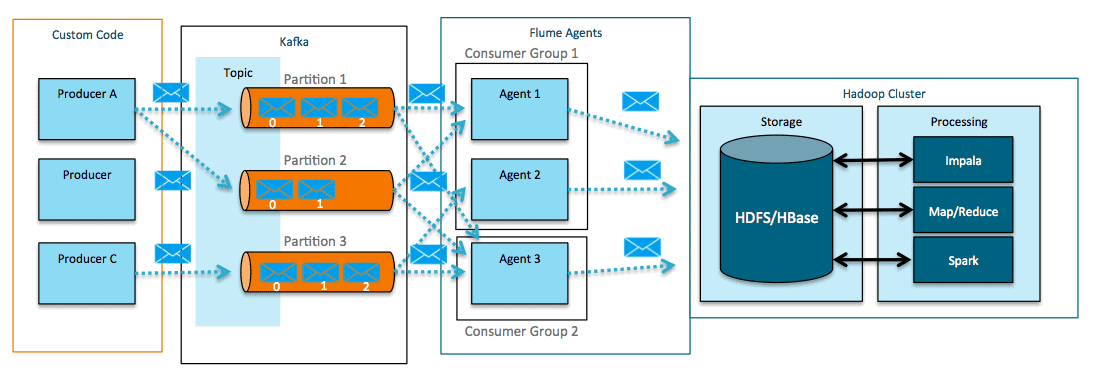
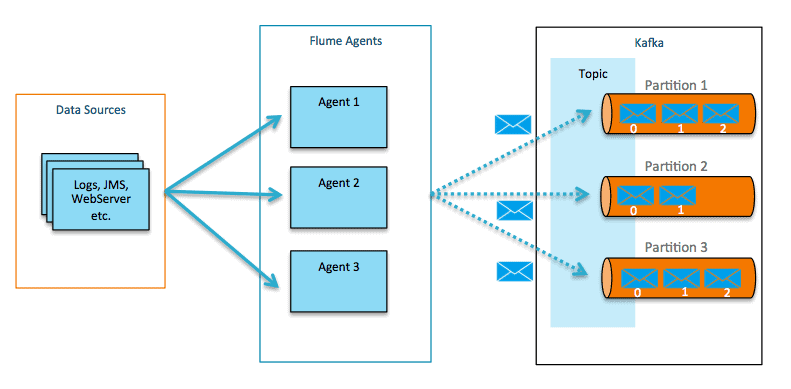
Big Data Challenge Round 2 Day 3 - **Apache Flume**

* Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of **log data**. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanism. It uses a simple extensible data model that allows for online analytic application
* Language: Java
* Problem with HDFS:
  + In HDFS, the file exists as a directory entry and the length of the file will be considered as zero till it is closed. For example, if a source if writing data into HDFS and the network was interrupted in the middle of the operation (without closing the file), then the data written in the file will be lost
  + Therefore, we need a reliable, configurable, and maintainable system to transfer the log data into HDFS
* Available Solutions:
  + **Facebook’s Scribe** – Scribe is an immensely popular tool that is used to aggregate and stream log data. It is designed to scale to a very large number of nodes and be robust to network and node failures
  + **Apache Kafka** – Kafka has been developed by Apache Software Foundation. It is an open-source message broker. Using Kafka, we can handle feeds with high-throughput and low-latency
  + **Apache Flume**
    - Efficiently stream data
    - Built for Hadoop scale
    - Always-on reliability
* Common Use Cases
  + Fraud detection
  + Internet of Things applications
  + Aggregation of sensor and machine data
  + Alerting/SIEM
* Apache Flume Meets Apache Kafka for Event Processing
  + Flume can act as both a consumer (above) and producer for Kafka (below)



Resources referred:

<https://blog.cloudera.com/flafka-apache-flume-meets-apache-kafka-for-event-processing/>

<https://en.m.wikipedia.org/wiki/Apache_Flume>

https://www.tutorialspoint.com/apache\_flume/data\_transfer\_in\_hadoop.htm