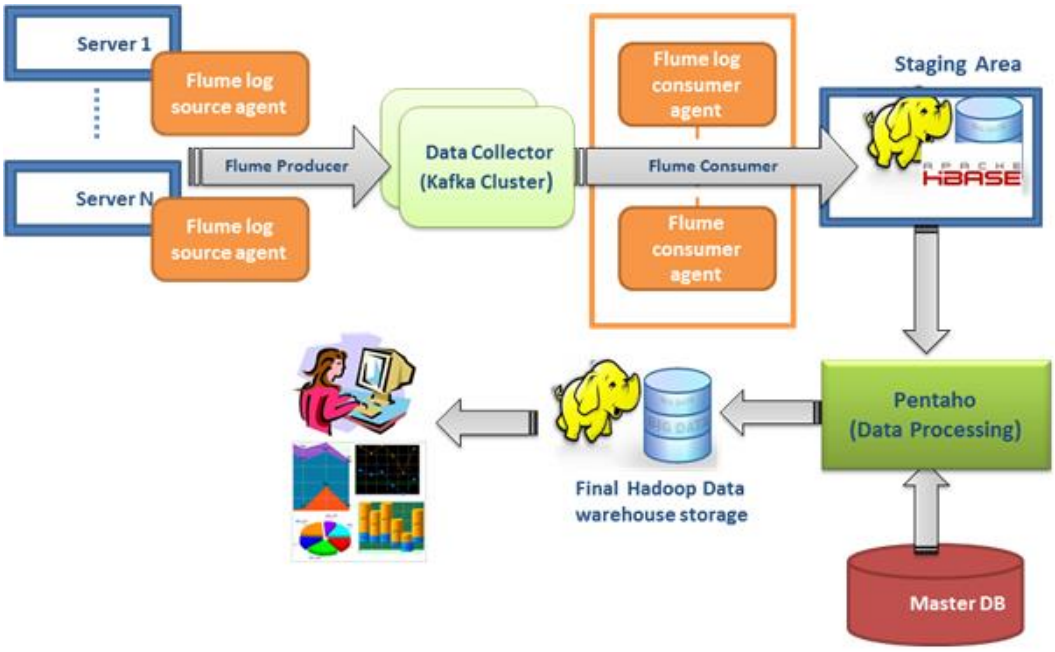
With our past experience, we recommend the indicative solution architecture shown in below diagram. The data ingestion flow can be achieved using tools like Flume, Filebeat or Kafka and the data processing can be achieved using tools like Pentaho, Spark streaming or Flume interceptor etc.



Solution Architecture Diagram

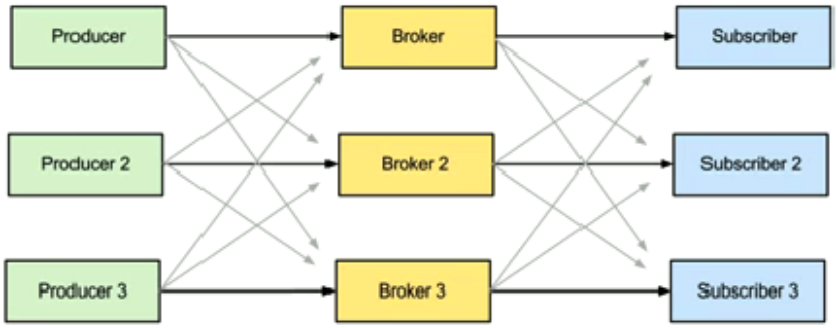
Below comes the description of each of the components, their duties and interactions with each other.

Flume agent on IPTV servers: The solution starts from collecting the logs messages from the servers. To collect log messages from servers, the flume agent will be installed on those servers. Flume agent is a lightweight java program which runs on JVM. This agent is configured to read the mentioned log file and push the log messages (`event` as per Flume's language) into the Data collector i.e. Kafka broker on configured batch size or time interval.

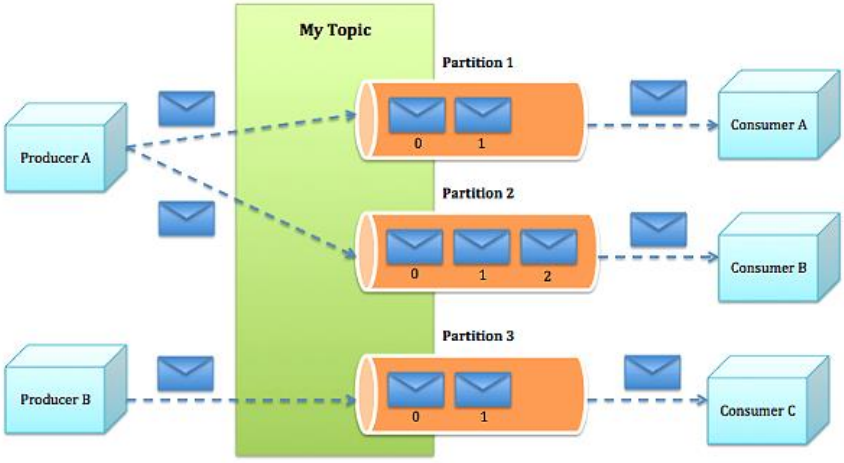
Data Collector/Kafka Broker: This module collects all the log messages pushed by the flume agents and keeps it safe for consumers to consume. In this design, Apache Kafka is chosen as the data collector because of its scalability, durability and low-latency advantages.

To collect the log messages, a topic T will be created in the Kafka broker and all the IPTV flume agents will logically push the log messages to that single topic. For high availability of the broker and for durability of the messages, we choose to have multi node-multi broker setup. To gain the optimal load balancing and true parallel processing ability of the multi broker setup, the topic T will be partitioned into n partitions using some key; where n >= number of subscribed consumers and each broker will contain one or more partitions. The

key can be subnet value, or a geo location value etc. The below diagrams show the discussed broker setup and partition concepts.



Multi Broker Setup with multiple producers and consumers



Topic and Partition Concept - shows parallel consuming ability

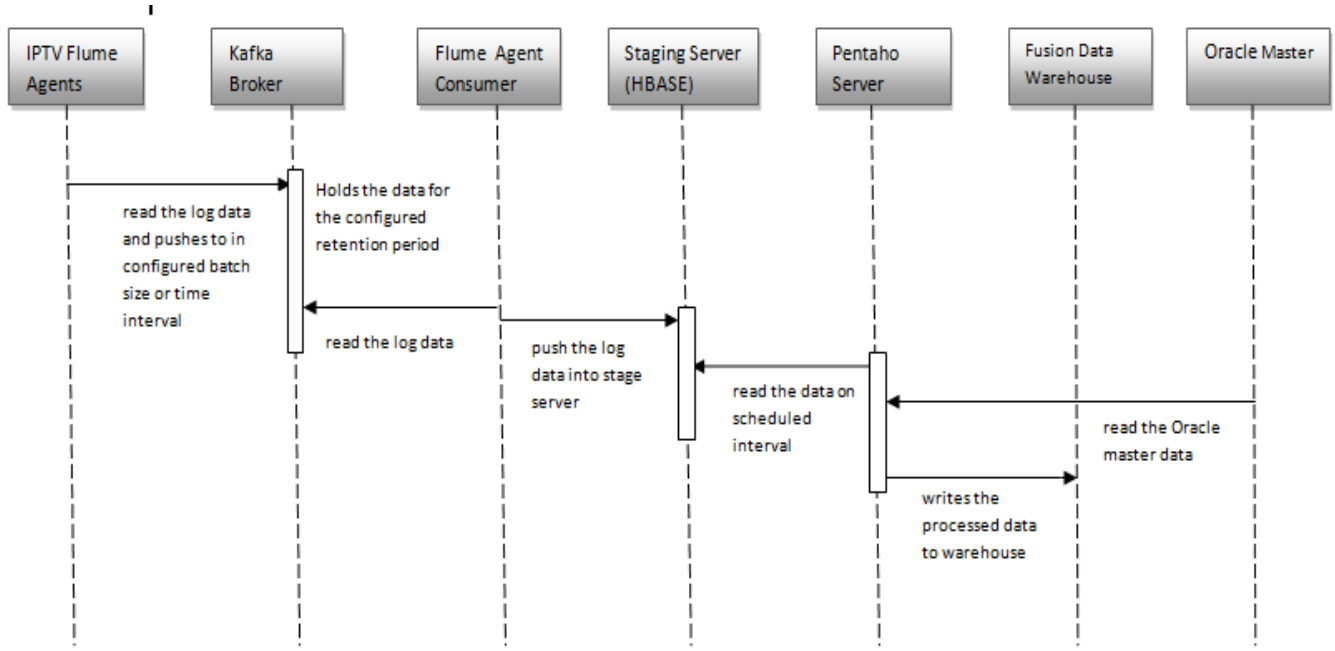
Flume Consumer: This layer consists of group of flume consumer agents connected to Kafka broker and subscribed to topic T. These agents together, will read the data from the subscribed topic and sinks into the staging area of HDFS layer. To achieve the parallel ability, all these agents will be grouped into one consumer group, so that these agents will decide among themselves who will read from which partition and will do load-balancing.

On other side these agents will be configured to sink the consumed log messages into HBase schema of the HDFS layer.

Pentaho ETL : Here is where the actual data processing happens. A scheduled ETL program will read the log data from HBase schema and also read the master content from the Oracle database, do the look up and processing and pushes the final data with defined structure into HDFS data warehouse layer.

IPTV Master Database: This master database holds the data of the IPTV customer and the media content details.

The below interaction diagram clearly states the log message flow and processing from end to end.



Interaction Diagram shows end to end log data processing flow

With the processed data being stored in HDFS data warehouse, it can be queried using a SQL interface like HIVE or can be exported into flat files such as txt, csv; which makes querying and utilization of data easier for business users.