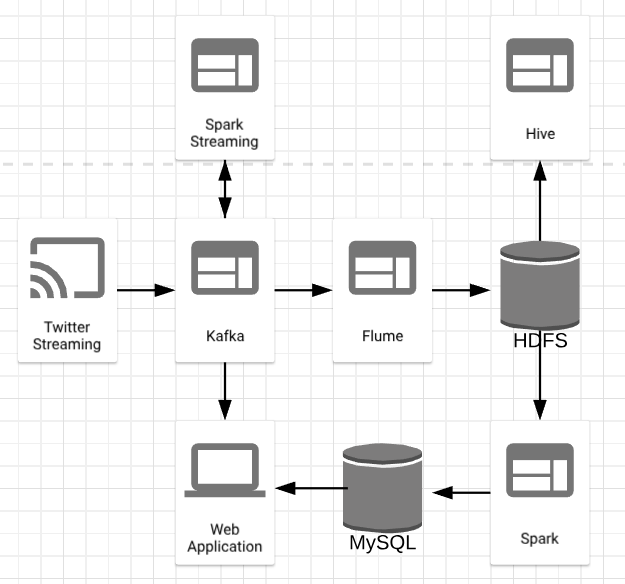
# **Real Time Twitter Analytics using Spark**

1. **Objective**

The project intends to use tweets to generate relevant insights into customers opinions about brands

1. **Features Built**
2. Real time counts of tweets
3. Collocations (sequence of words that co-occur more often than would be expected by chance)
4. **Technical Data Flow Diagram**



1. **Technical Data Flow Details**
2. Node Js is used to call Twitter Streaming API
3. Tweets are then streamed into Kafka topic “TwitterStream” to be consumed by other applications
4. Spark Streaming subscribes to the Kafka topic “TwitterStream”, process the data in real time and puts back the data into another kafka topic “Counts”. The data is then broadcasted to the web application using Socket.io
5. Flume subscribes to the kafka topic “TwitterStream” and writes the data to HDFS
6. Hive external tables are built on top of HDFS to allow for querying the data
7. Spark batch process is run on the data that is stored in HDFS to generate collocations using the mean variance approach
8. The web application is built on Node Js
9. **Technology Stack**
10. Spark
11. HDFS
12. Hive
13. Flume
14. Kafka
15. MySQL
16. Node Js