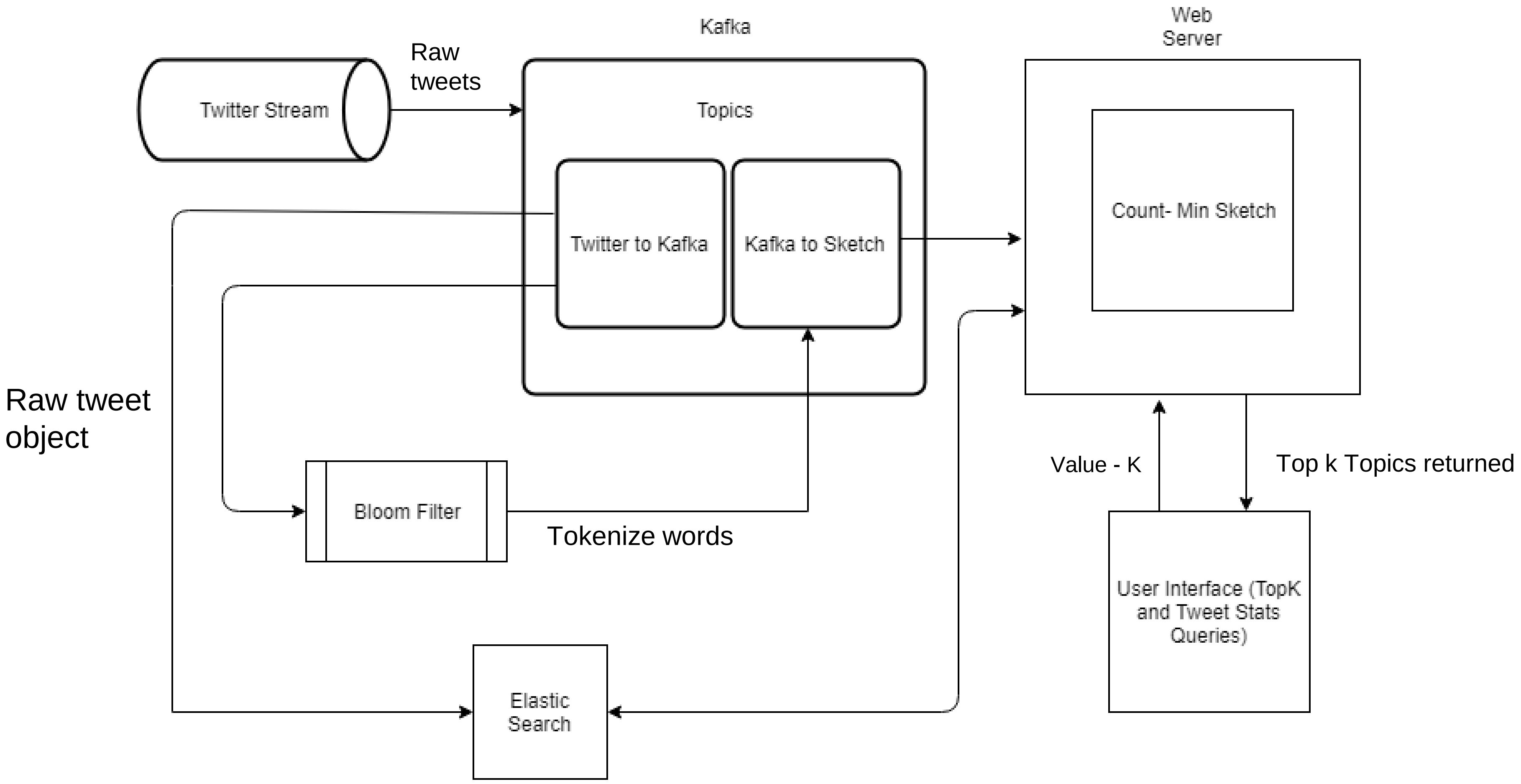
Topic: Trending Topic Miner

Team

Harsh Patel
Jaymin Desai
Shrijeet Joshi

Data Pipeline



Use Case

- Get top k words/topics using count-min sketch from a stream of tweets and query the topics into elastic search to get tweets containing those words/topics.
- These top k topics would be a trending word or topic as we are using live twitter stream data.
- User gets the trending tweets on a webpage.

Kafka

- Kafka acts as a buffer for the streaming data.
- Two topics:
- **“twitter2kafka”** topic to consume twitter data and send tweets to elastic search
- Use the same topic to send data to bloom filter and return tokenized values to “**kafka2sketch”** topic.

Elastic Search

- Dumping tweets into specified index from Kafka
- Supports fast querying over large corpus of text.
- Top-K words returned by the count min sketch shall be queried to get trending tweets.
- We query tweets based on keyword from Elastic Search.

Count-Min Sketch

- Sub-linear space data structure
- Supports sub-logarithmic and constant time complexity querying.
- Returns count of a particular element added in the sketch.
- Works well on streaming data as it is fine if we get false positives.

Bloom Filters

- Space Efficient Data Structure
- Using it to check for stop words
- Used NLTK Stop and other words like http, rt, etc.
- Check the membership of a stop word in the data structure
- Ignore the stop words and proceed to tokenize the string.

Python-Flask

- Used to implement a stateful container microservice.
- Container to persist the count-min sketch data structure.
- Hosts the micro-service which gets hit by our User Interface

User Interface

- Webpage to display the trending tweets.
- Enter the value of K on the webpage.
- Hits the custom micro-service end point and returns a JSON object
- Consumes the JSON object returned and displays tweets.
- This is the final part of our end to end solution.

Future Scope

- Implementation of other queries on the count-min sketch data structure.
- Providing support for live querying for region based tweets to get information pertaining to a particular region.
- Leverage other DRPC queries for various practical use cases.