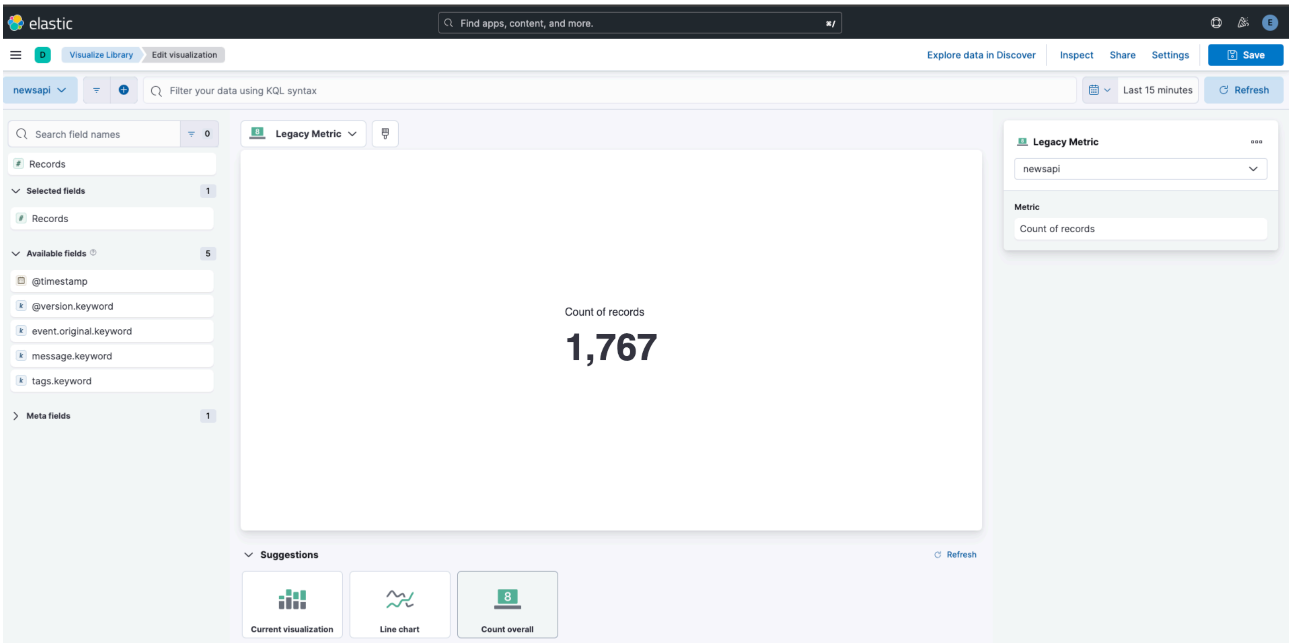
Assignment 3
Part 1
Report

The real-time text data source that I have used for this part is NewsApi. I First generated a key and started fetching data every 60s and storing to Kafka topic1. Then I performed spark structured streaming where the data form Kafka topic1 is fetched, processed the data to extract named entities(tokanize, stopwords etc). Then find the count of each named entity and store the result to Kafka topic2. Configured Elasticsearch, Kibana and then by configuring the Logstash established a path to transfer data in Kafka topic2 to Elasticsearch.

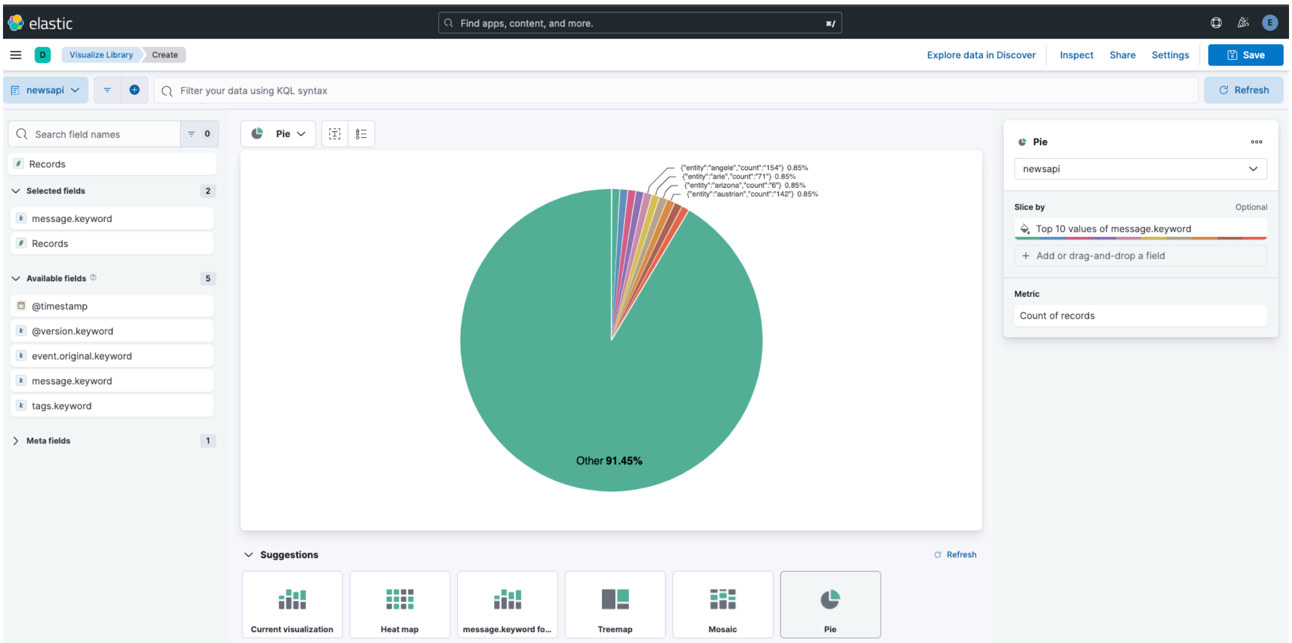
Finally done visualisation and the resultant screen shots are below:



Count of records
1,767

Legacy Metric
newsapi
Metric
Count of records

The above image shows the total count of the records. For every time frame the total count will be increasing.



Pie
newsapi
Slice by Optional
Top 10 values of message.keyword
Add or drag-and-drop a field
Metric
Count of records

The above plot diagram represents the top 10 named entities with highest counts. Each colour represents each named entity and the green colour portion is the total remaining counts.