1. Spark Streaming with Real Time Data and Kafka and ELK stack

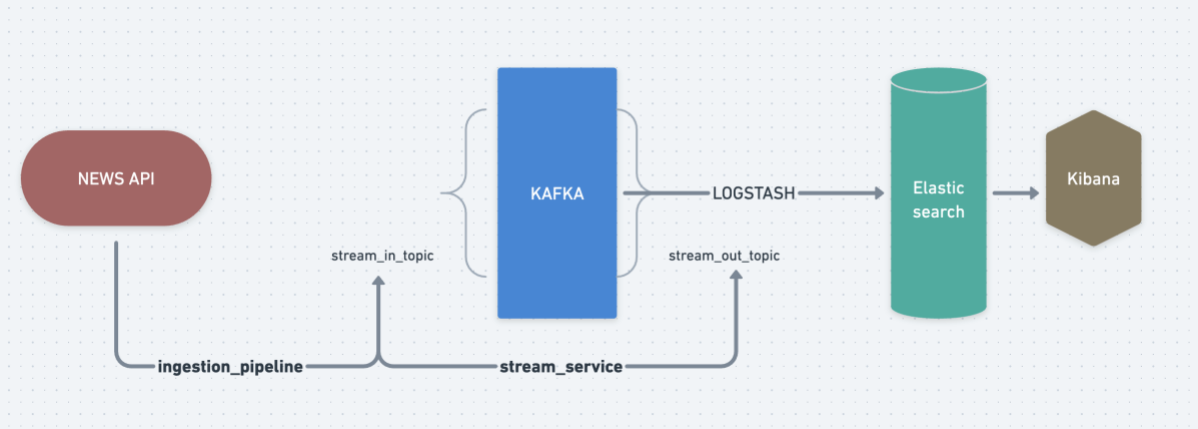
We are performing 'Named Entity Recognition (NER) count' on the live streaming data from newsapi [https://newsapi.org/].

Data source -

https://newsapi.org/docs/endpoints/everything

This API searches through millions of articles from over 80,000 large and small news sources and blogs. This endpoint suits article discovery and analysis.

Below is the pipeline architecture –



Application configuration -

Project configurations are defined in file **app-config.yml** and can be changed as per need –

```
bootstrap_servers: kafka:29092
input_topic: stream_in_topic
output_topic: stream_out_topic
api_key: f3abe15c03e04488b3ade0a1e5eb1167
news_api_url:
https://newsapi.org/v2/everything?q=the&apiKey={api_key}&sortBy=publishedAt&l
anguage=en
checkpoint_location: '/tmp/checkpoint'
news_api_request_delay: 20 # seconds
```

Running steps –

The project is setup in docker-containerized environment and can be easily launched using '**docker-compose up'**.

It will automatically download all the base images and will launch the necessary services. Docker daemon has to be running for it. Please see here or more details - https://docs.docker.com/compose/

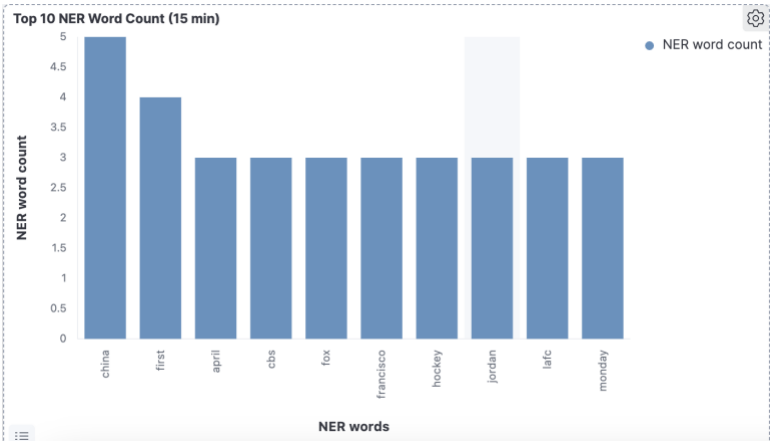
Respective launched applications can be found at –

- Kafka - <localhost:9092> [Inside containers it will be accessible as <kafka:29092>]
- Elasticsearch - http://localhost:9200/
- Kibana - http://localhost:5601/ [*Please import [**'export.ndjson'**] file in kibana to import dashboard analysis.]*

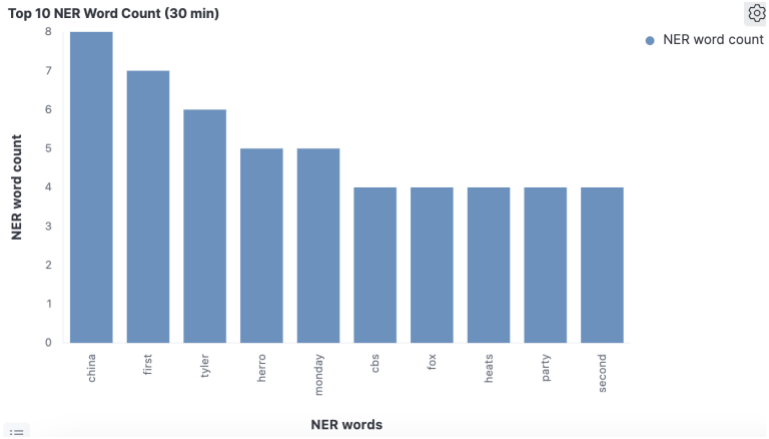
Results –

NER word count analysis has been demonstrated using various plots. Here is the description for them –

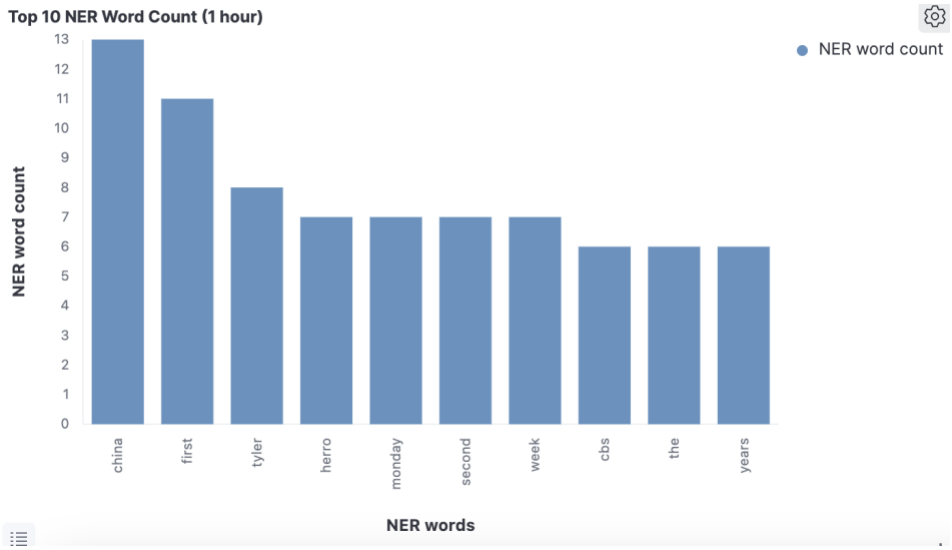
1. **Top 10 NER Word Count (All time)**



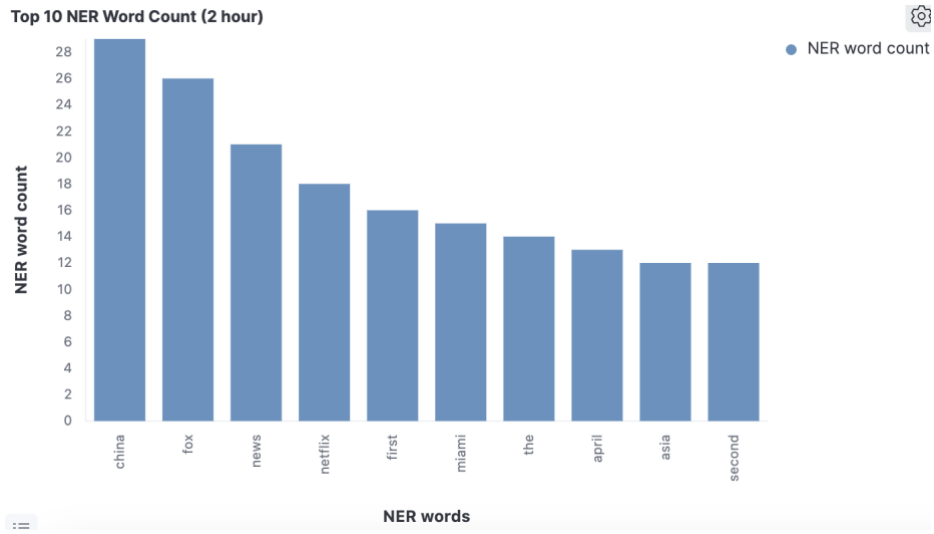
Top 10 NER Word Count (15 min)



Top 10 NER Word Count (30 min)



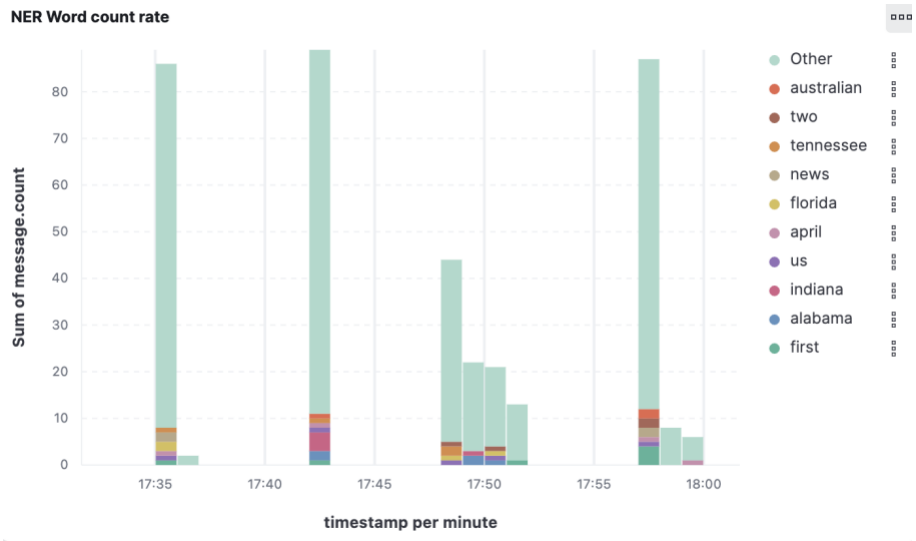
Top 10 NER Word Count (1 hour)



Top 10 NER Word Count (2 hour)

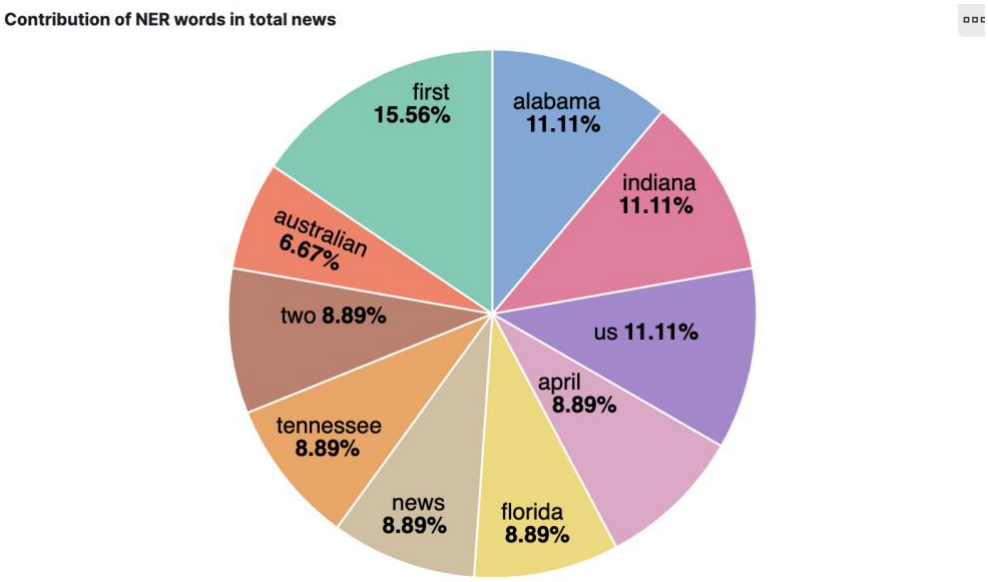
The above graph shows the top 10 NER words that have occurred most frequently in all time.

2. NER Word count rate



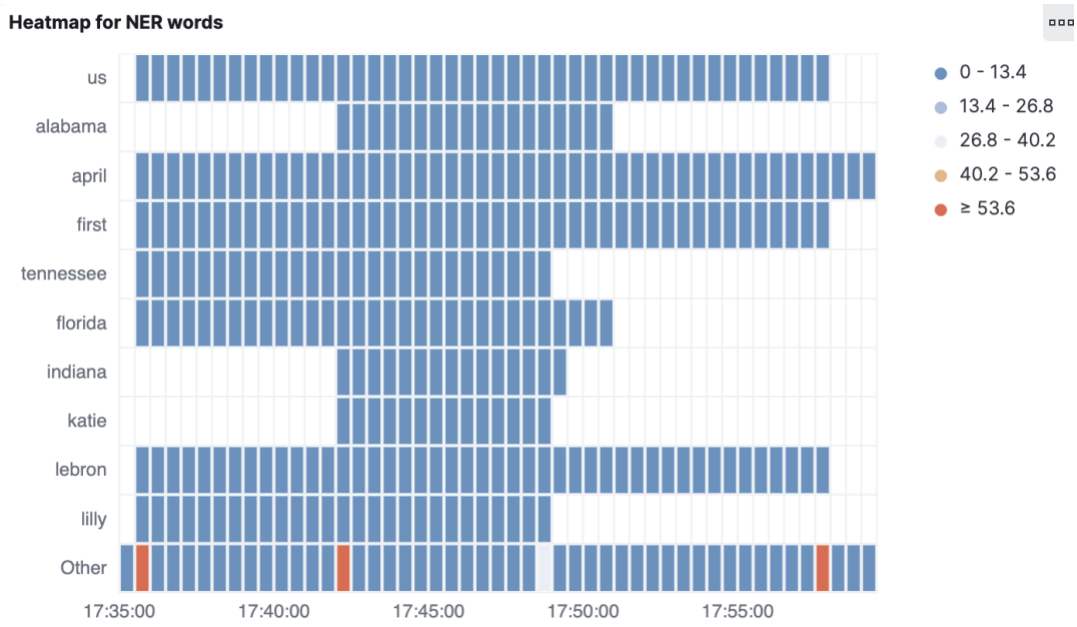
The stacked barplot graph shows the rate of NER words every minute.

3. Contribution of NER words in total news



The pie chart shows the individual contribution of top 10 NER words in all time news.

4. Heatmap for NER words



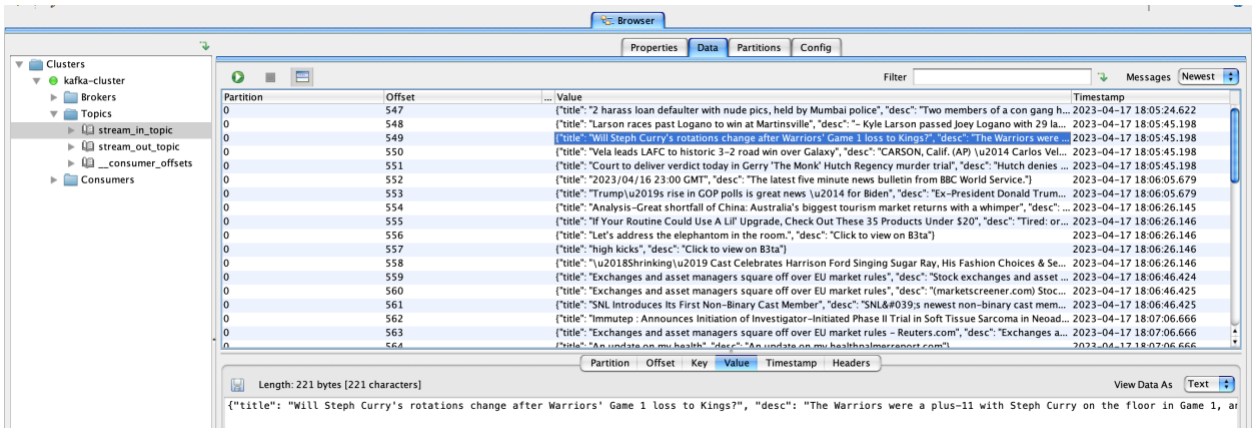
The heatmap shows the frequency of NER words at various timestamps. Heat is represented using color scales.

Result Analysis

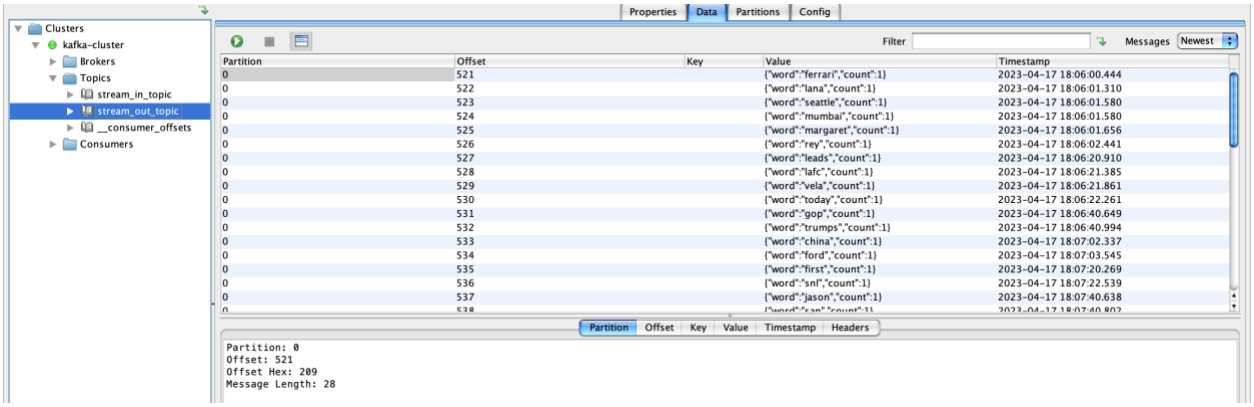
The graphs shows the running frequency count of different NER words extracted from real time news. We observer that NER words like ‘US’, ‘florida’, ‘indiana’, ‘katie’ have been observed in the data at different times.

Here are further screenshots of data at various stages –

1. At input Kafka topic – (stream_in_topic)



2. At output Kafka topic – (stream_out_topic)



3. Inside Elasticsearch –