Real-Time Named Entity Recognition Streaming Pipeline

Overview

This project establishes a real-time streaming pipeline designed to extract and visualize the most frequently mentioned named entities from live news articles. The components used include NewsAPI for fetching real-time news, Apache Kafka for message streaming, Apache Spark Structured Streaming for Named Entity Recognition (NER), Logstash to ship data into Elasticsearch, Elasticsearch for data storage, and Kibana for visualizing the results.

Data Source

The data for this pipeline is sourced from NewsAPI, which provides real-time news articles. Instead of fetching only the top headlines, our producer queries random keywords across different topics (such as technology, sports, business, health, and politics) using the `/v2/everything` endpoint. This approach ensures a more diverse range of news articles entering the pipeline.

Pipeline Components

- **Producer**: A Python application that fetches fresh news articles every minute and sends them to the Kafka topic `topic1`.
- **Spark Structured Streaming Application**: Reads the articles from Kafka topic `topic1`, uses SpaCy to perform Named Entity Recognition (NER) on the text, counts the occurrences of named entities, and writes these counts to Kafka topic `topic2`.
- **Logstash**: Consumes the data from Kafka topic `topic2`, parses the JSON entity counts, and ships them into Elasticsearch.
- **Elasticsearch**: Stores the named entities along with their counts for efficient querying.
- **Kibana**: Provides visualization of the top 10 most frequently mentioned entities over time.

Results and Observations

We observed the evolution of top named entities over different time intervals, with the following key insights:

Time Interval	Observations
After 15 minutes	Entities such as "KSL-news", "1-year-old", and "10" were dominant, primarily from tech and political news.
After 30 minutes	The data began to show more sports and health-related entities as news topics rotated.
After 45 minutes	Education and business-related entities emerged more prominently.
After 60 minutes	A well-balanced set of entities from multiple topics was visible.

The changes in entities reflect the dynamic nature of incoming live data, as the pipeline adapts to various news topics fetched randomly.

Challenges and Improvements

- **Duplicate News**: Initially, the news fetched every minute was repeating. This was resolved by switching from the `/v2/top-headlines` endpoint to `/v2/everything` and adding random keyword searches to fetch more diverse articles.
- **Kibana Visualization**: We refined the Kibana dashboard to display the top 10 entities clearly, without causing confusion due to breakdowns in the data.

Notes

- The News Producer fetches articles related to random topics such as technology, sports, health, politics, and more every minute.

- Spark performs Named Entity Recognition on the news text, counts how often each entity is mentioned, and streams the data to Kafka.

- Logstash processes the data from Kafka and sends it to Elasticsearch for storage.

- Kibana is used to visualize the top 10 named entities over time, providing valuable insights into the most frequently mentioned topics in real-time.