CS 6350: Big Data Management and Analytics

Assignment 3

Names of students in your group:

1. Kushal Choudhary (kxc240000)

Number of free late days used: 0

Part 1: Spark Streaming with Kafka

Code: https://github.com/Kushal3121/1. Spark Streaming with Kafka

1. Data Source

The data for this project was obtained from the **NewsAPI**, a live REST-based feed providing real-time headlines from major global news outlets.

The Python-based producer script (newsapi_producer.py) fetches new articles periodically from the API using the endpoint https://newsapi.org/v2/top-headlines.

Each news record contains fields such as:

- title - headline text
- author - article author (if available)
- source.name - news outlet
- publishedAt - timestamp of publication

These headlines serve as continuous text input for Named Entity Recognition (NER). The producer publishes each news headline as a JSON message into the **Kafka topic topic1**, simulating a live data stream.

2. Processing and Visualization

A PySpark Structured Streaming job (stream_ner_to_counts.py) reads live messages from Kafka (topic1) and applies SpaCy's NER model (en_core_web_sm) to detect entities such as:

- Organizations (e.g., Apple, Microsoft)
- Persons (e.g., Elon Musk, Joe Biden)
- Locations (e.g., United States, China)

The Spark job maintains a running count of how many times each entity appears and writes these results to Kafka topic topic2.

Logstash then consumes messages from topic2, parses them as JSON, and indexes the entity counts into Elasticsearch under the index name ner-entities.

Finally, Kibana visualizes these results:

- A bar chart displays the top 10 most frequent entities at any given time.
- A line chart tracks entity frequency changes over time (5-minute intervals).

3. Results and Interpretation

The visualization dashboards show that certain entities, such as “Apple,” “AI,” “Trump,” and “United States,” consistently appear among the most mentioned.

This pattern indicates their strong presence in global news discussions and trending topics during the streaming period.

The bar chart highlights static entity dominance at a single point in time, while the line chart demonstrates temporal variation, some entities rise or fall in mentions as new headlines arrive.

Overall, the results confirm:

- Successful real-time ingestion and NER extraction from live sources.
- Proper aggregation and indexing in Elasticsearch.
- Continuous visual updates in Kibana, verifying an end-to-end streaming analytics pipeline.

4. Conclusion

This project demonstrates how Apache Kafka, Spark Streaming, and the Elastic Stack can be integrated to perform real-time NLP analytics.

By automating entity extraction from live news, it provides insights into which people, companies, or countries dominate current news cycles, enabling scalable and continuously updating text analytics.

5. Appendix: Screenshots

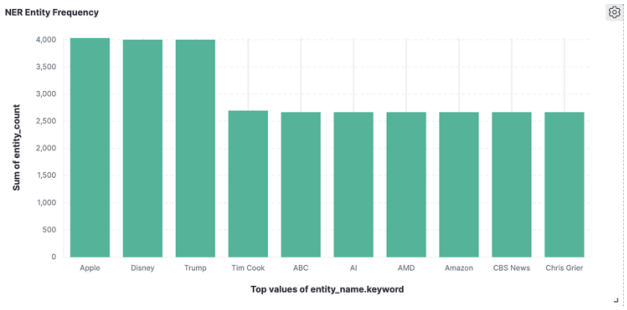

Fig. 1. Top 10 Entities (15 min)

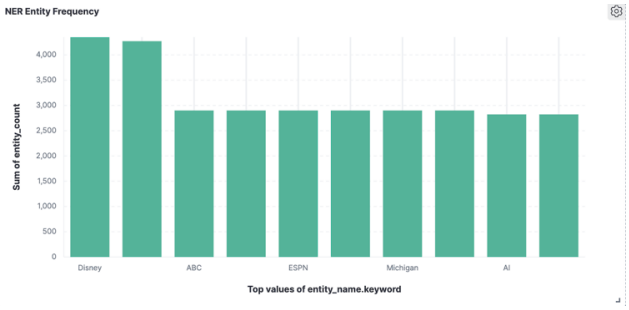

Fig. 2. Top 10 Entities (30 min)

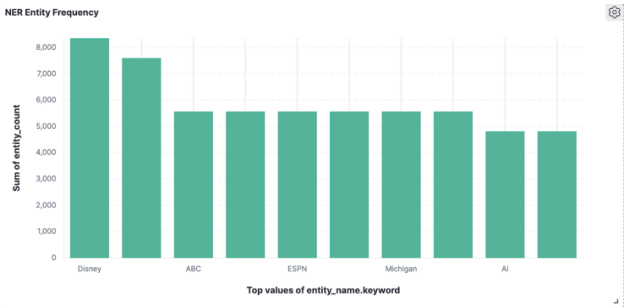

Fig. 3. Top 10 Entities (45 min)

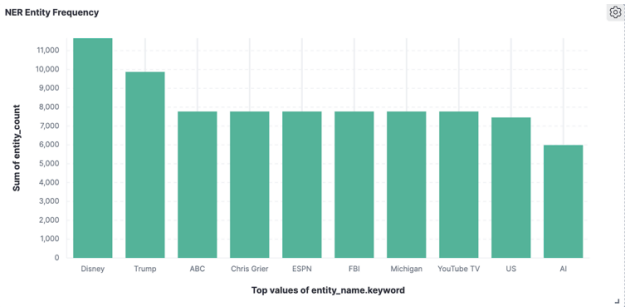

Fig. 4. Top 10 Entities (60 min)