Introduction

This project focuses on implementing a data processing workflow using Reddit as the data source, specifically targeting the 'movies' subreddit. Real-time comments are streamed using PRAW (Python Reddit API Wrapper) in adherence to the project requirements. The objective is to perform named entity recognition, count occurrences, and visualize the data through an integrated stack of technologies.

1. **Data Source and API**:
   The project leverages Reddit as the data source, with a focus on streaming live comments from the 'movies' subreddit. PRAW facilitates efficient access to Reddit's API, allowing seamless real-time data collection.

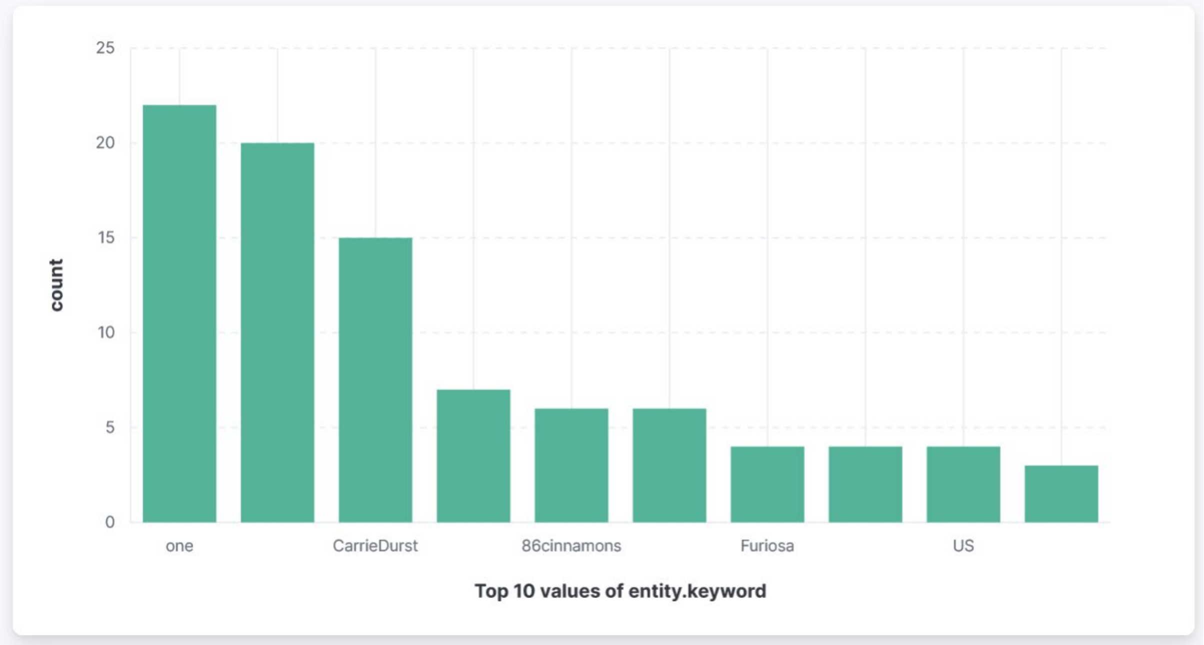
2. **System Configuration**:
   In line with project requirements, incoming data is serialized into JSON format and sent to a Kafka topic named topic1. This setup ensures a robust and scalable data pipeline capable of handling high-volume data streams.

3. **Data Processing**:
   A PySpark application processes the data streamed from topic1. It extracts and counts named entities in the comments, providing insights into topics and entities frequently discussed in the 'movies' subreddit community.

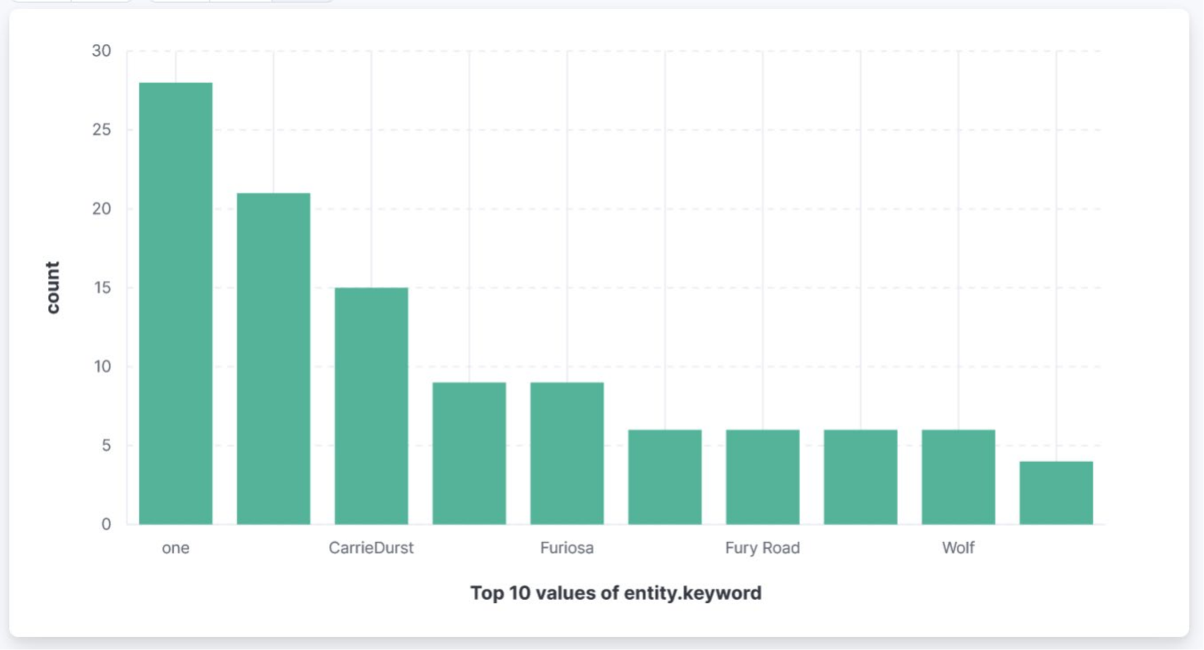
4. **Data Transmission and Visualization**:
   Processed data is forwarded to another Kafka topic, topic2, as per the project specifications. Logstash then transfers this data to Elasticsearch, enabling visualization in Kibana. The result is a bar chart highlighting the most frequently mentioned named entities, offering a clear view of community discussions.

Analysis of Top 10 Named Entities

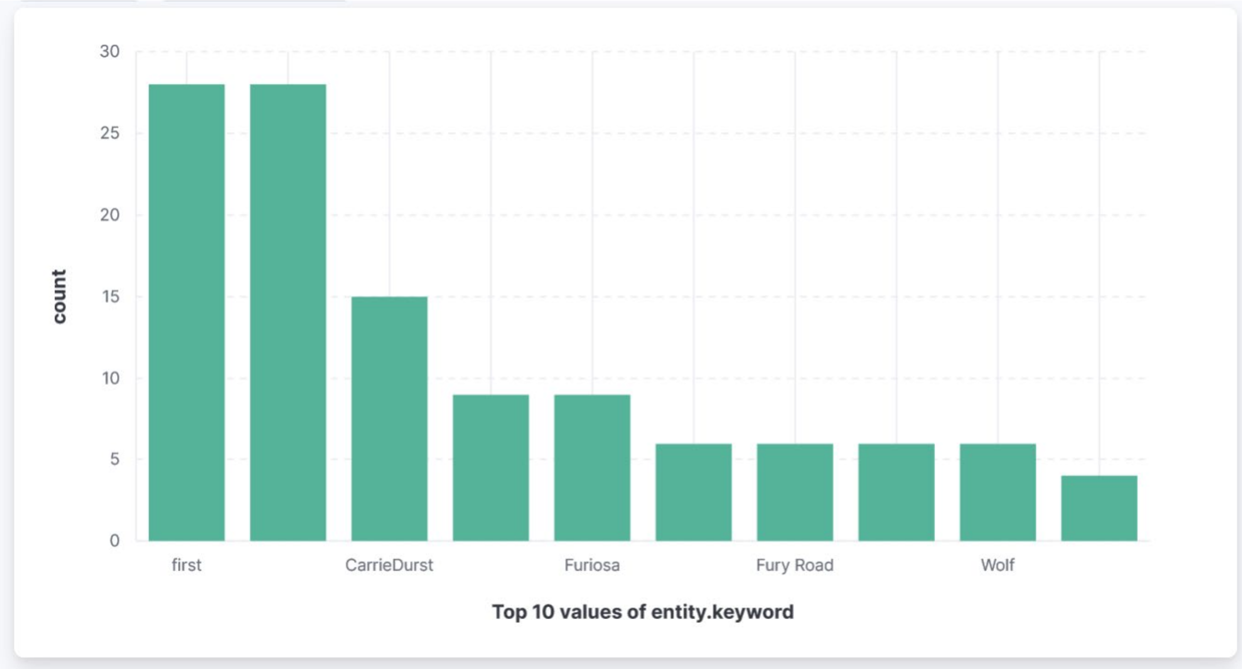
1. **Visualization at 15 minutes**



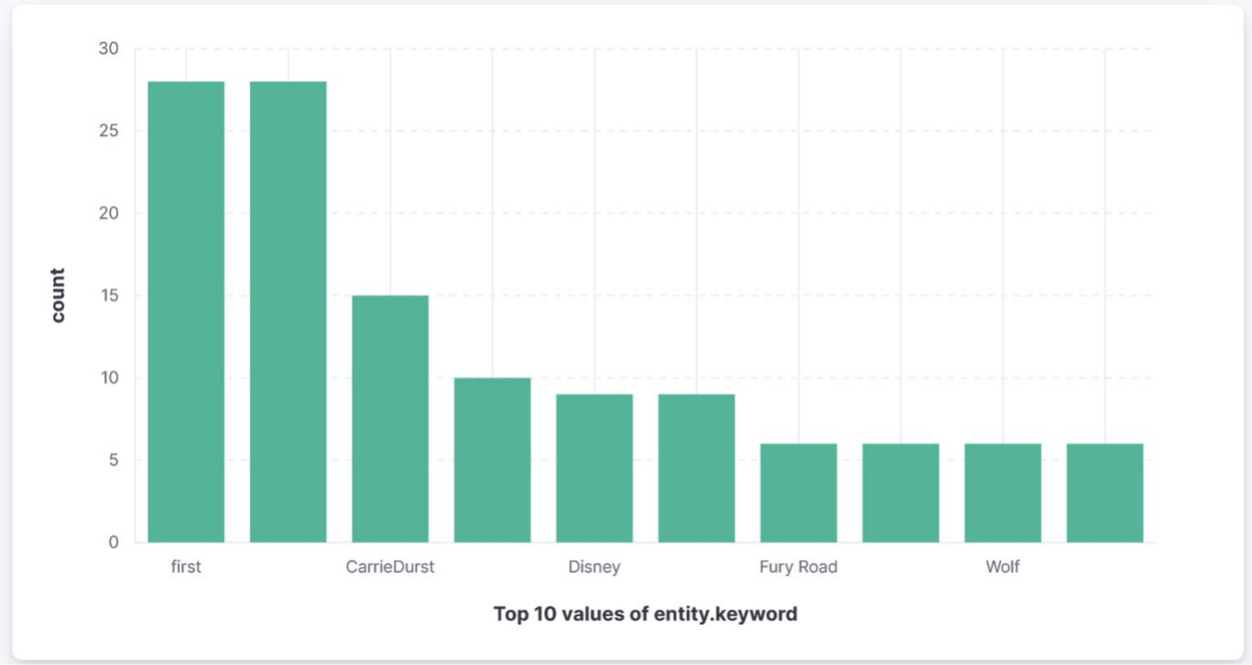
2. **Visualization at 30 minutes**



3. **Visualization at 45 minutes**



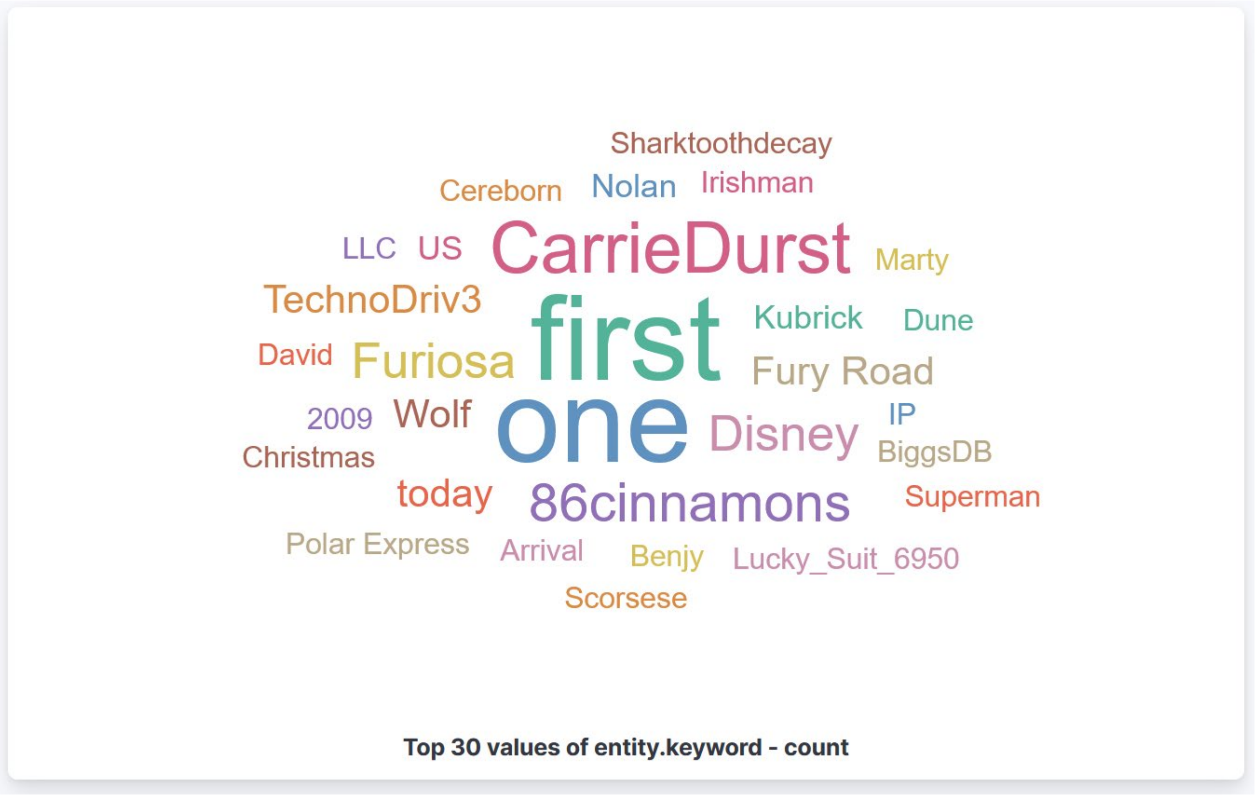
4. **Visualization at 60 minutes**



Observation of Named Entities in Movies discussion

Our analysis focused on identifying commonly mentioned movies, directors, and topics using named entity recognition (NER). The results were displayed in Kibana as a bar chart, showcasing the most talked-about entities in the community.

From the data we collected, users frequently discussed directors like Martin Scorsese and George Miller, films such as Furiosa, and their thoughts on acting and storytelling. These findings reflect the community's enthusiasm for both timeless classics and recent releases. The insights provide a snapshot of what's trending among movie lovers and can be useful for anyone interested in understanding public opinions or tracking film-related discussions.


Sharktoothdecay
Cereborn
Nolan
Irishman
LLC
US
CarrieDurst
Marty
TechnoDriv3
first
Kubrick
Dune
David
Furiosa
Fury Road
Wolf
one
Disney
IP
Christmas
BiggsDB
today
86cinnamons
Superman
Polar Express
Arrival
Benjy
Lucky_Suit_6950
Scorsese
Top 30 values of entity.keyword - count