Introduction

This project focuses on implementing a specified data processing workflow using the Reddit platform as the data source. We adhere closely to the outlined project requirements by streaming real-time comments from the 'nba' subreddit using PRAW (Python Reddit API Wrapper). The goal is to analyze these comments for named entity recognition, count occurrences, and visualize the data using a series of interconnected technologies.

1. **Data Source and API**: The data is sourced from Reddit, specifically the **'nba'** subreddit, which provides a continuous stream of user comments. We use PRAW for its efficient access to Reddit's API, enabling real-time data streaming.

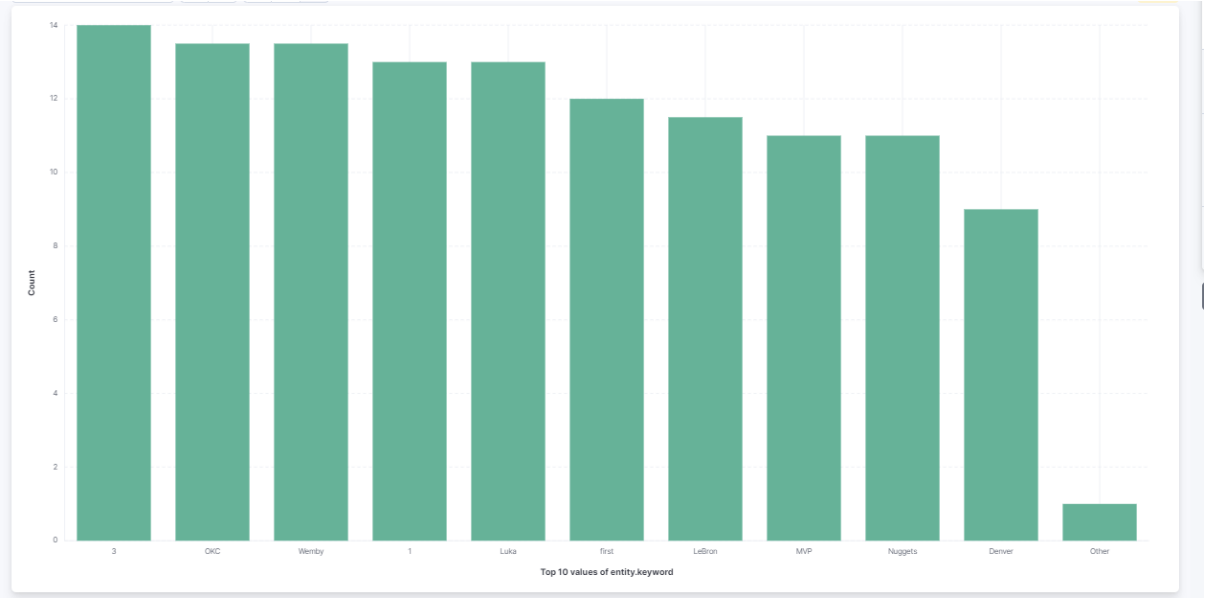
2. **System Configuration**: As per the project specifications, we serialize incoming data into JSON and send it to a Kafka topic named 'topic1'. This approach ensures that our data pipeline is robust and prepared for high-volume data flows.

3. **Data Processing**: Following the project instructions, our PySpark application processes the stream from 'topic1'. It identifies and counts named entities within the comments, aligning with the task of understanding community focus within the NBA discussions.

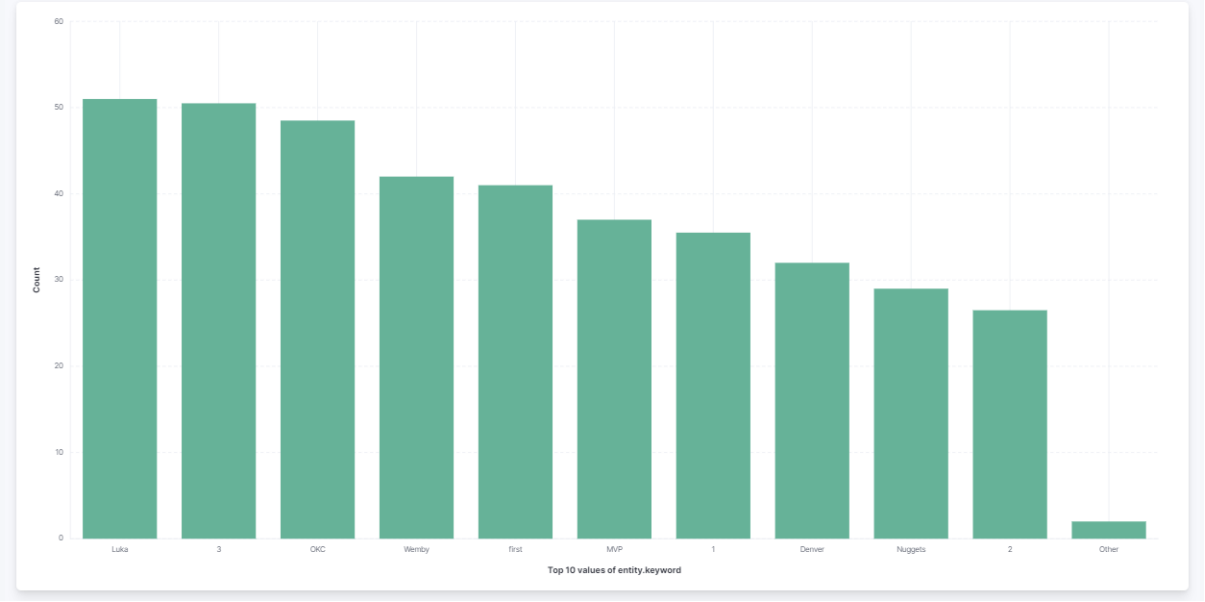
4. **Data Transmission and Visualization**: In line with project guidelines, we transfer the processed data to another Kafka topic, 'topic2'. From there, Logstash forwards this data to Elasticsearch, which we then visualize in Kibana as a bar plot of the most frequently mentioned named entities.

Analysis of Top 10 Named Entities

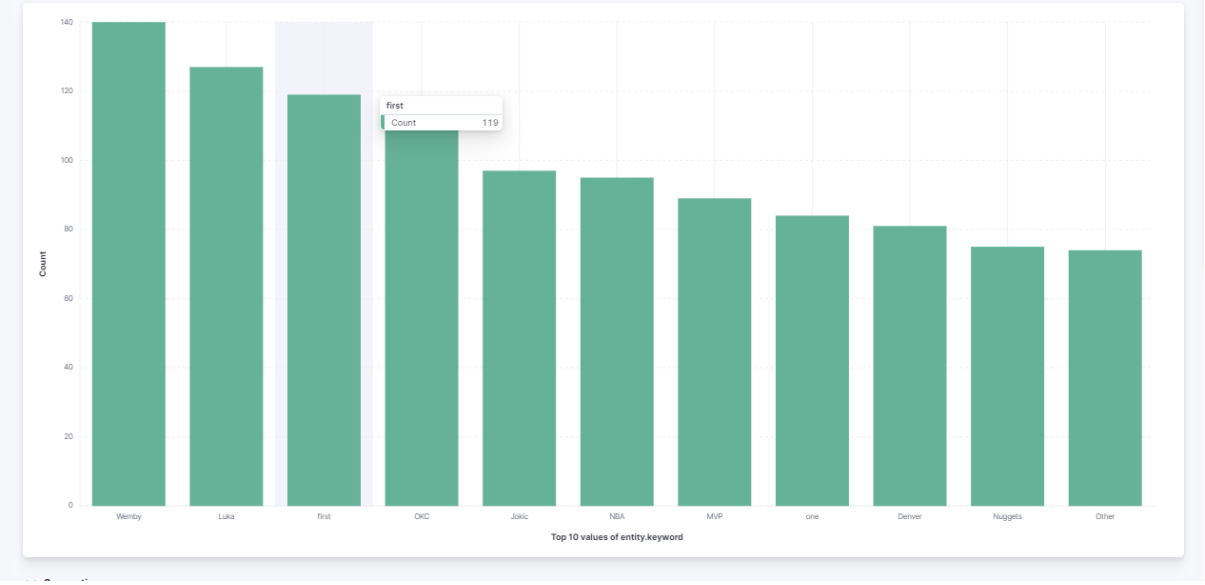
1. **Initial Visualization at 10 Minutes**



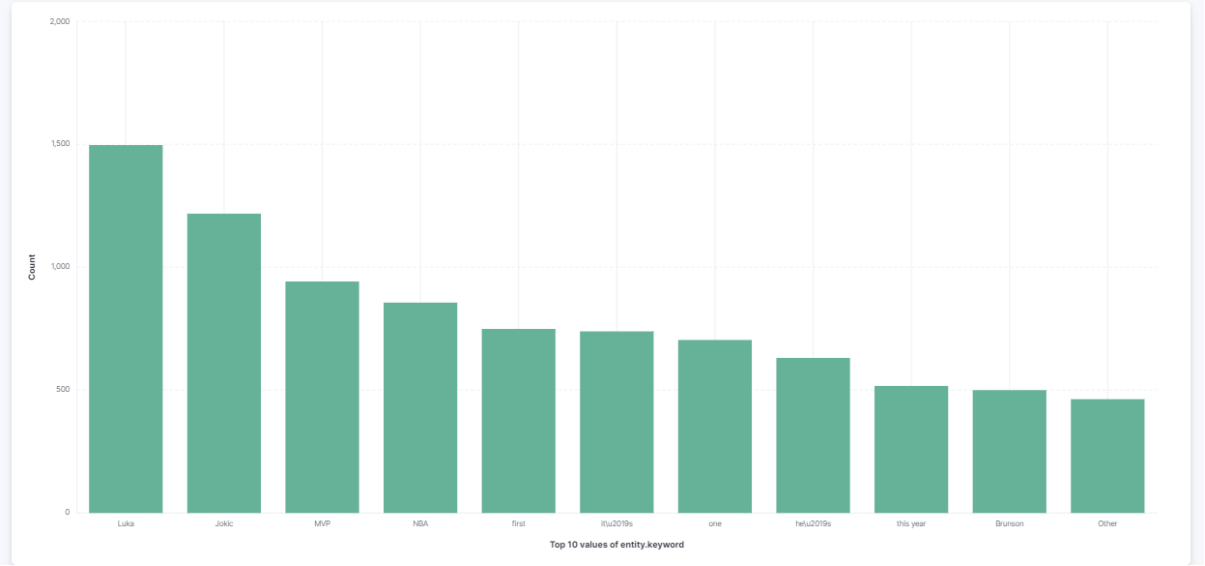
2. **Updated Visualization at 30 Minutes**



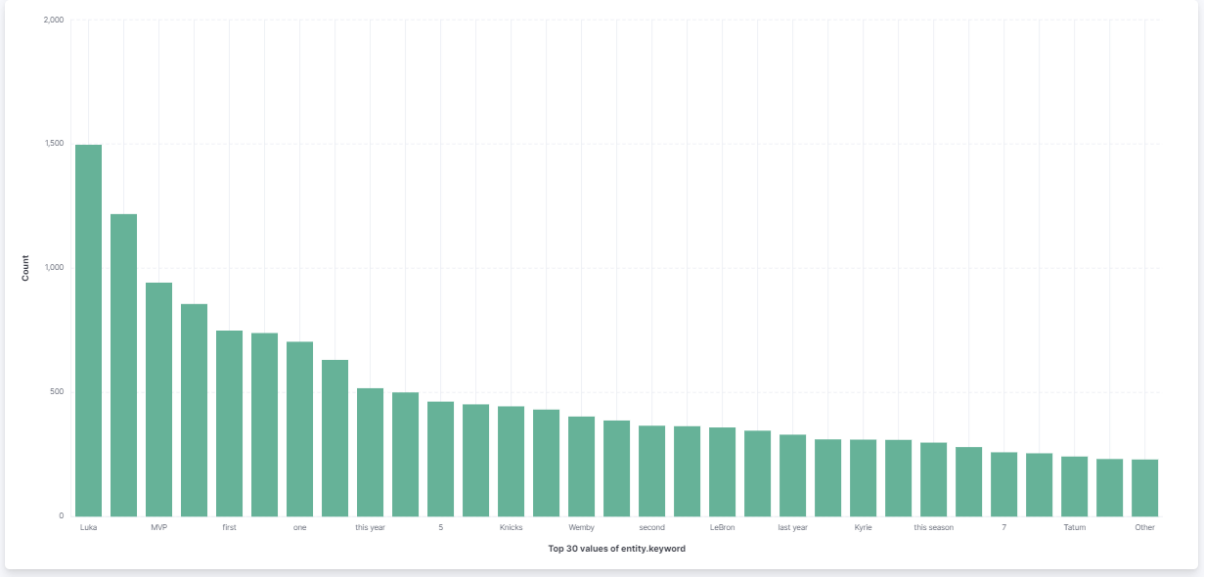
3. Further Visualization at 45 Minutes



4. Final Visualization at Two Hours



Top 30 Named Entities



Word Cloud Graph of Named Entities

Observation of Named Entities in NBA Discussions

In our analysis of real-time comments from the 'nba' subreddit, the entities "Wemby," "Luka," "OKC," "MVP," "NBA," and "Jokic" stood out as the most frequently discussed topics. "Wemby" and "Luka" reflect individual player popularity and performance, while "OKC" suggests strong team-specific interest. "MVP" discussions are indicative of ongoing debates about player achievements, and "NBA" serves as the overarching theme for all discussions. "Jokic" highlights focus on specific player impact and league status. These trends suggest a vibrant mix of player, team, and league-wide discussions within the community.