Social Media Analysis using Kafka

Surmeet Kaur Jhajj - 54245972
Aiswarya Manikandan- 94646410
Rohan Rajeev - 38249388

Data Streaming in today's world

- Large amounts of data being generated from social media today
- Instantaneous input and fast analysis is required
- Frameworks consists of: i) distributed data ingestion; ii) distributed data processing; iii) data visualization
- Apache Kafka is a distributed streaming platform for messaging
- It follows the publish/subscribe system
- Data from multiple sources is sent to a Producer which streams it to the consumer for further processing

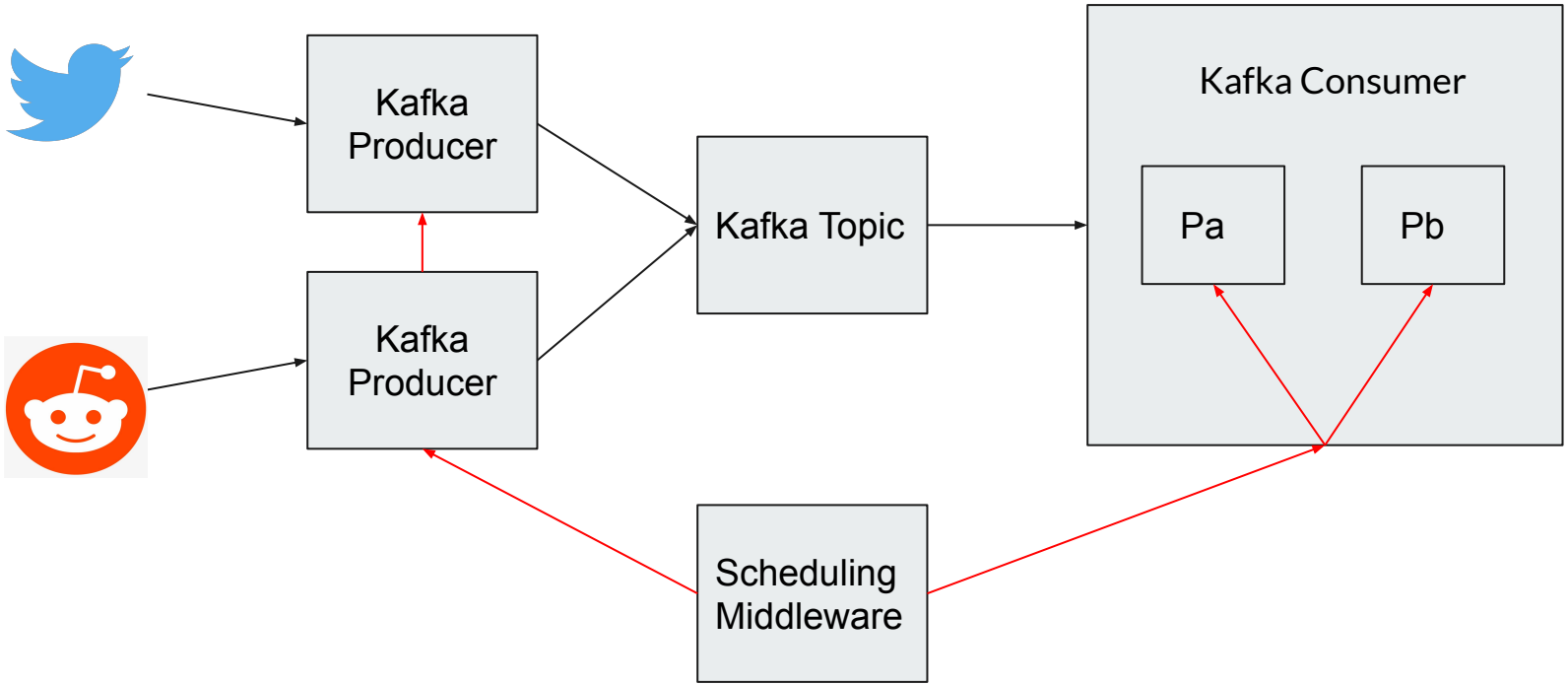
Our Objective

- Establish a data-pipeline with ingestions from multiple sources
- Define a scheduling middleware which determines the selection of processing logic based on the data ingestion rate

What we propose....

- Data is live Streamed from two social Medias : Twitter and Reddit
- This Data is fed into the respective Kafka Producers
- Producers, with the help of the scheduler, stream this data to the consumer.
- The scheduler is responsible for normalizing the input rates, acting accordingly in cases of high input.
- Consumer further processes this data according to processing requirements.

Architecture



References

[1] L Magnoni, “Modern Messaging For Distributed Systems”, Journal of Physics, Conference Series

[2] Rajiv Ranjan, “Streaming Big Data Processing in Datacenter Clouds”, IEEE Cloud Computing

(Volume: 1 , Issue: 1 , May 2014)

[3] Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabrizi, “Developing a Real-time Data Analytics Framework For Twitter Streaming Data”, 2017 IEEE 6th International Congress on Big Data

[4] Hassan Nazeer, Waheed Iqbal, Fawaz Bokhari, Faisal Bukhari, “Real-time Text Analytics Pipeline Using Open-source Big Data Tools”