Big data platform final project executive summary

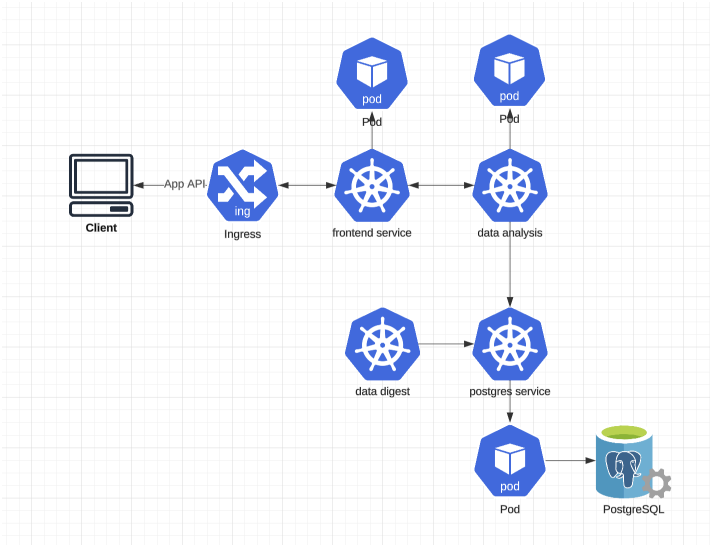
Project overview:
We have developed a kubernetes based application which is designed to store tweets in a database and query this database in order to find meaningful insight into the data.

Data description
We have a CSV file with the tweets of the 20 most popular twitter users.
The dataset includes fields such as the tweets author, tweets content, and the data when the tweet was published. We also created a parsed content column which is a product of cleaning the content string from URL and special characters and parsing the string to array. We save this data into tweets_df.csv. Before data insertion, we explode this table by each string in the parsed content column to create a table partitioned by word. This table is being used to pull insight about the content easier

K8 cluster overview and rationale
Our project uses a microservices architecture where each component is encapsulated in its own pod. We have the following 3 separate pods:

- A postgres database that stores our data persistently.
- Backend that fills the posgres tables with data
- A front end which allows the user to enter the tweet id they wish to perform sentiment analysis on.
- A back end which interacts with our database and performs the sentiment analysis. [data_analysis service]

The design rationale behind this configuration is to allow for easy scaling of resources based on demand to a specific service. Deploying each component separately allowed us to work better as a team as each team member could focus on their own portion of work without heavily relying on other parts of code. This is because each microservice can be developed independently of the other.



Challenges faced and how we overcame them

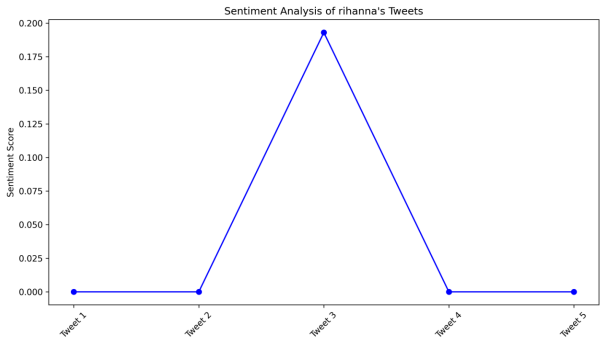
While developing and testing the pod which would hold the database everytime we ran the pod it would crash. This presented a problem which at first we struggled to solve. We then used the kubectl logs command in order to dive into the problem and understand what was causing the pod to crash. We realized our code which set out the schema for our database was incorrect. After finding this bug by going through the logs we were able to change the lines of code causing it and redeploy the pod with a fixed database image.

Key insights from the data

sentiment analysis:

After running sentiment analysis on the tweets we found that many of them had negative sentiment. This confirmed out hypothesis that most people on twitter tweet about negative things rather than positive

Below is a graph showing this for the tweets of Rihanna. As you can see all tweets have a positive sentiment score less than 0.2 out of 1. Clearly she is mostly tweeting about not positive things.



Top 10 Users by Content Volume:

{
“TheEllenShow", 3147, "jimmyfallon", 3123, "ArianaGrande", 3104, "YouTube", 3077, "KimKardashian", 2939, "katyperry", 2924
, "selenagomez", 2913, "rihanna", 2877, "BarackObama", 2863, "britneyspears", 2776
}

Top Tweets by Likes and Shares

Selenagomez: "My heart is absolutely broken. I miss you Christina https://t.co/KWGwZZlj4t"

Analysis Based on Token: token=’hate’

"most_liked_tweet": ["Truth is last thing we need right now is hate, in any form",
"selenagomez", 157981
],
"most_shared_tweet": [
"I hate broke bitches",
"rihanna", 78359
],
"year_most_mentioned": [
"2016", 88] ----> **Black Live Matters protests on Year 2016** }