Nikunj Malpani
INFO - I 535

Big Data for Football Analysis

Introduction

In the world of football, data analytics has become one of the most significant ways to optimize in-game processes and performance. To continuously improve player performance and reduce the risk of injury, clubs implement programs that use big data analytics to monitor player performance during games and after practice sessions. During the upcoming football World Cup tournament in Qatar, players will have a more evidence-based way to argue for time on the pitch. With billions of followers worldwide and millions of tweets every minute, every hour, we will have enough data to understand what people are more interested in and what is in the trend.

It's the latest incursion of numbers into the beautiful game. Data analysis now helps to steer everything from player transfers and the intensity of training to targeting opponents and recommending the best direction to kick the ball at any point on the pitch. There is a lot of data collected every day from all the tweets around the world, and it's critical to extract value from it so that relevant analyses can be performed and meaningful solutions and decisions can be taken. In this project, we will analyze the trends and patterns of tweets to show what is trending and how can we create a pipeline to make this analysis an end-to-end procedure, and develop some forecasting models for future instances of tweets.

Background

Football is a beautiful game as half the world's population watches it and with World Cup around the corner, this number increases significantly. It is a very tough game and it's critical to win on and off the pitch for any team to really win the world cup, as there are many mind games at play, and with Twitter being one of the biggest social media platforms, it is imperative we understand what is the world talking about, does it affect the team and their playing strategy, does it affect any one particular player more than another and finally work on something to create a pipeline so that every time we click refresh, we get the answer as to what is being trended and if this can affect the game's outcome in any manner.

The word daily in itself says that this is an ever-increasing amount of data and out of those very few data available today, this problem satisfies all 5 Vs of big data. The data relating to the various aspects of Football, for example, the total number of tweets today, is very fast-moving, since half the world's population is watching, thus by its very nature it has velocity, it has a lot of volumes since it is reported in huge quantities. The data also has a variety in terms of features that are reported, for example, languages and format. It is also very important to check for the trustworthiness of the data source and also the truthfulness of the data thus even veracity is addressed here. Lastly, we derive value through all this data and make appropriate analysis for

Nikunj Malpani
INFO - I 535

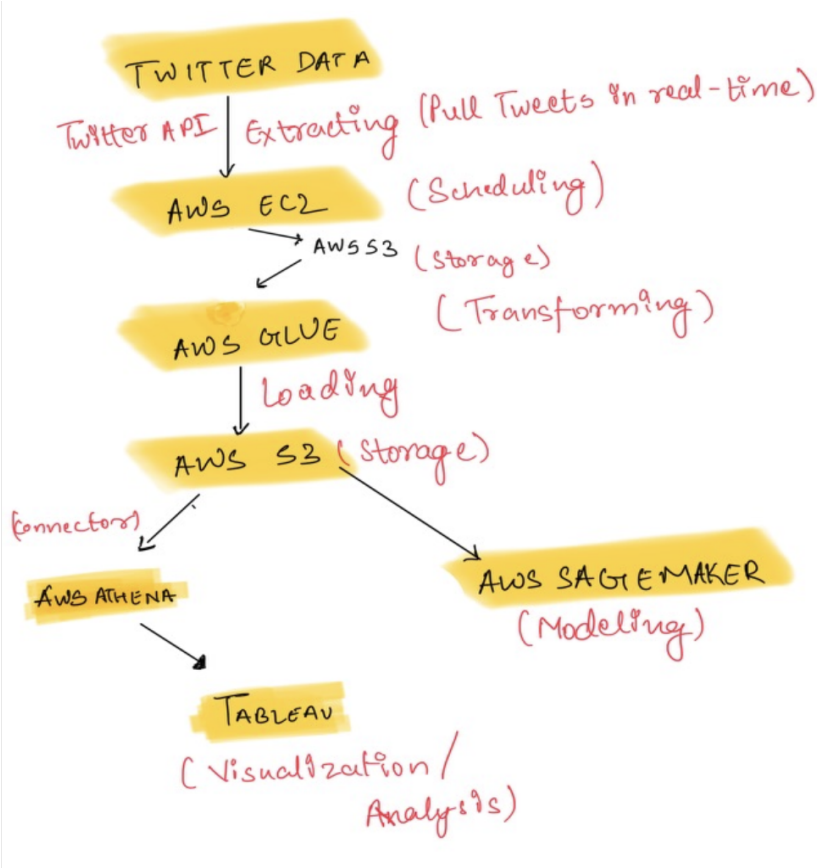
Methodology

For the following problem we will use AWS cloud services as a base project image instance and we will code all our logic into the services of AWS.

AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals. With AWS Data Pipeline, you can regularly access your data where it's stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR.

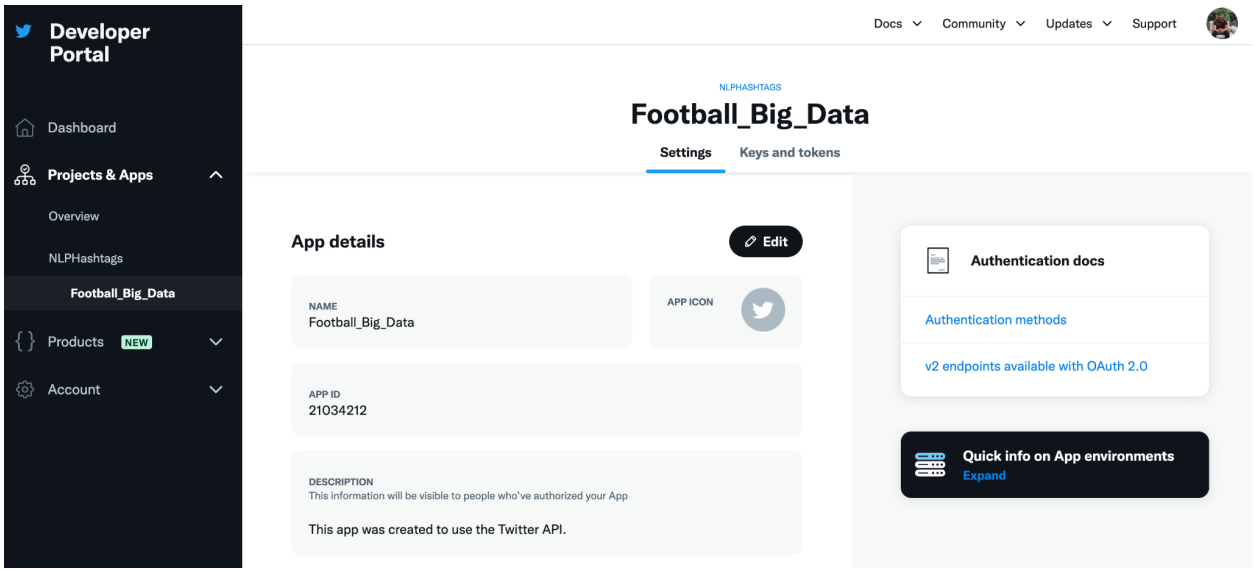
AWS Data Pipeline helps you easily create complex data processing workloads that are fault tolerant, repeatable, and highly available. You don't have to worry about ensuring resource availability, managing inter-task dependencies, retrying transient failures or timeouts in individual tasks, or creating a failure notification system. AWS Data Pipeline also allows you to move and process data that was previously locked up in on-premises data silos.

PIPELINE BREAKDOWN



Nikunj Malpani
INFO - I 535

- A **Twitter** user account is a requirement to have a **"developer" account**, to get the access keys to the tweets and the information that we need to pull from the twitter.



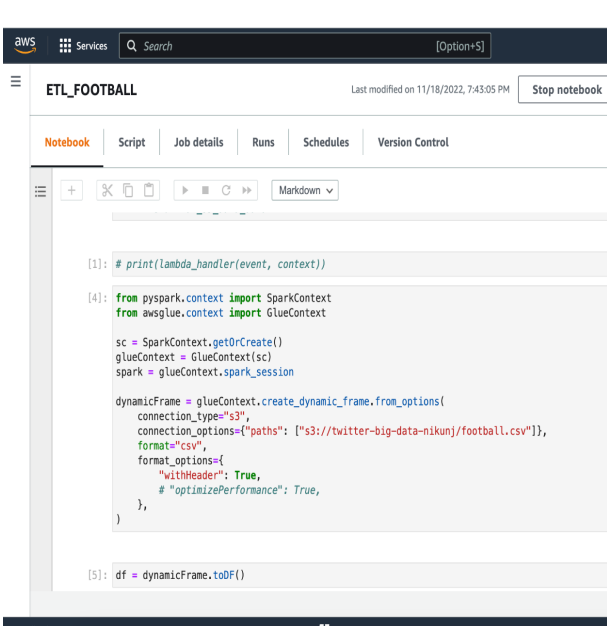
- After that, A python script on **AWS EC2** runs and pulls tweets with our keyword in real-time. The script writes the tweets into an AWS S3 bucket in the csv format as we wrote in the script. We are going to run the script on an AWS EC2 instance. Steps below:

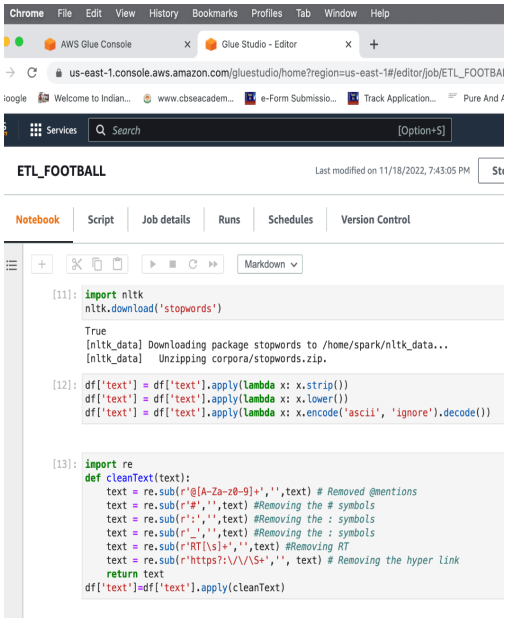
```
# connect to your AWS EC2 instance either through the browser or SSH # terminal

# change to your project directory
cd /path/to-your-project

# create a virtual environment
pip3 virtualenv
virtualenv venv

# activate your virtual environment
source venv/bin/activate

# install the libraries
pip3 install -r requirements.txt

# If you get errors on the dotenv file, use the following command
pip3 install python-dotenv

# run the script continuously in the background
screen

# run the script
python3 pull-tweets.py

# exit the screen
Ctrl+A then PRESS D

# THE SCREEN IS NOW EXITED BUT THE SCRIPT IS RUNNING IN THE BACKGROUND
```

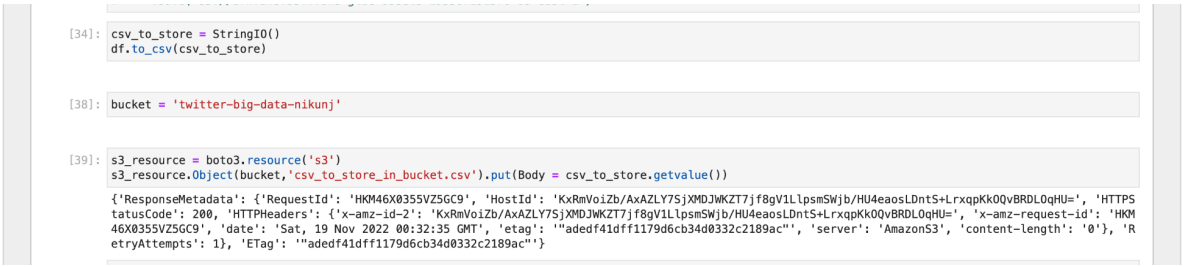
```
def scrape(words, date_since, numtweet):
        # Creating DataFrame using pandas
    db = pd.DataFrame(columns=[
                                    'description',
                                    'location',
                                    'retweetcount',
                                    'text',
                                     'date',
                                    'hashtags'])

        # We are using .Cursor() to search
        # through twitter for the required tweets.
        # The number of tweets can be
        # restricted using .items(number of tweets)
    tweets = tweepy.Cursor(api.search,
                               words, lang="en",
                               since_id=date_since,
                            #until=date_until,
                               tweet_mode='extended').items(numtweet)


        # .Cursor() returns an iterable object. Each item in
        # the iterator has various attributes
        # that you can access to
        # get information about each tweet
    list_tweets = [tweet for tweet in tweets]

        # Counter to maintain Tweet Count
    i = 1
```

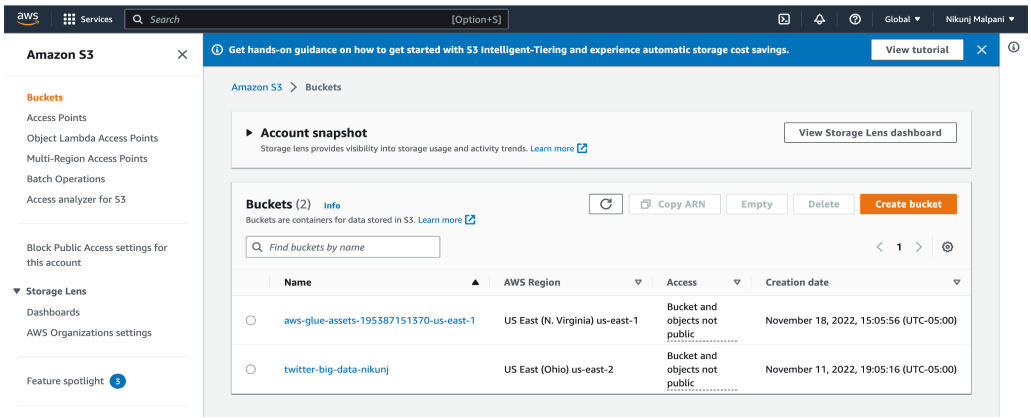
```
    # list for extracting information about each tweet
for tweet in list_tweets[::-1]:
    description = tweet.user.description
    location = tweet.user.location
    retweetcount = tweet.retweet_count
    date=tweet.created_at
    hashtags = tweet.entities['hashtags']


            # Retweets can be distinguished by
            # a retweeted_status attribute,
            # in case it is an invalid reference,
            # except block will be executed
    try:
        text = tweet.retweeted_status.full_text
    except AttributeError:
        text = tweet.full_text
    hashtext = list()
    for j in range(0, len(hashtags)):
        hashtext.append(hashtags[j]['text'])


            # Here we are appending all the
            # extracted information in the DataFrame
    ith_tweet = [description,location,retweetcount, text,date, hashtext]
    db.loc[len(db)] = ith_tweet


            # Function call to print tweet data on screen

    # printtweetdata(i, ith_tweet)
    i = i+1
filename = 'football.csv'
```

- **Amazon Glue** extracts the data from S3 bucket and performs the transformations as mentioned in the Glue notebook and then reloads the new data into the S3 bucket. Here we also use Pyspark to load the data into the Glue as even though the dataset right now is small but in future if we want to scale it, then Pyspark would be a better choice as it allows you to process your data efficiently in a distributed fashion. It is the most sought after big data processing platform, providing capability to process data on the petabyte scale. PySpark gives us the flexibility to read data in various formats like csv, parquet, json or from databases.









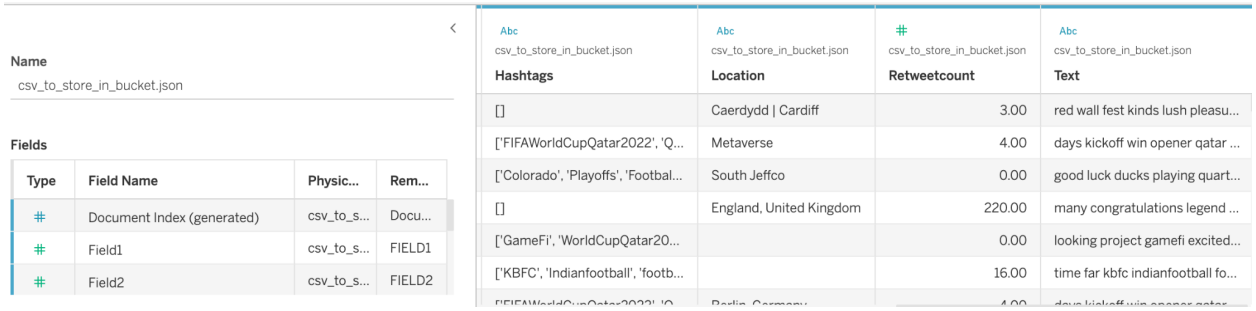
- Then we connect our file that is saved in S3 to Tableau using **AWS Athena** connector. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

- We will extract the data from the S3 using AWS Athena connector and put it in **Tableau** to make our data visualizations. Tableau is an effective data visualization and a Business Intelligence tool with a very easy interface to get visualizations.

We can understand the high level overview of the methodology from the image below where the given methodology is a nice blend of extraction, transformation and loading (ETL) pipeline all in one jupyter notebook and it uses the power of parallel processing framework *PySpark* and the advantages of AWS services which makes it so easy. The methodology also makes use of the powerful Machine Learning technique of Natural Language Procesing and the use of an effective visualization tool Tableau.

Results

After loading the dataset, we first look at the basic analysis of the tweet counts and number of retweets with respect to the hashtags and combinations of them.



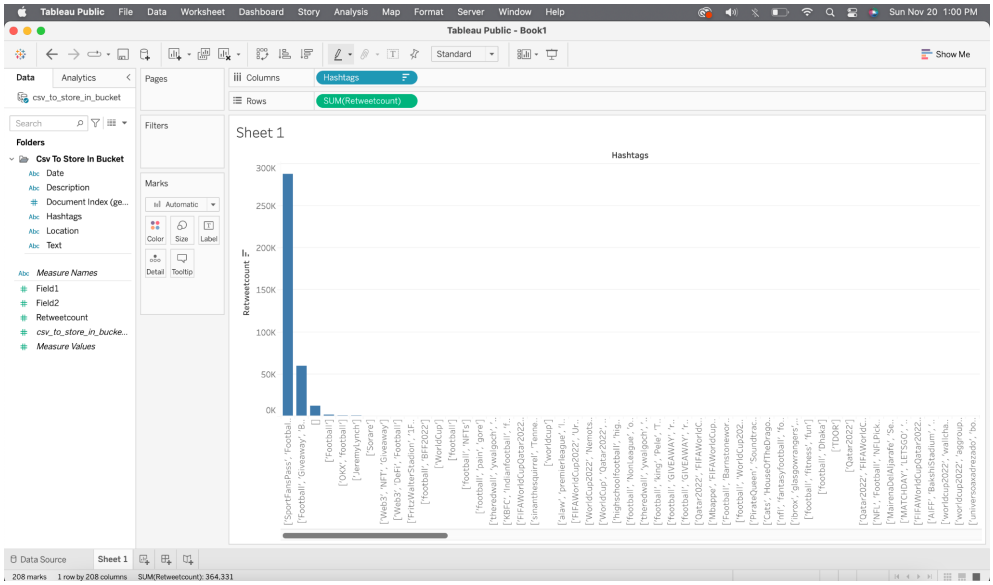
Name

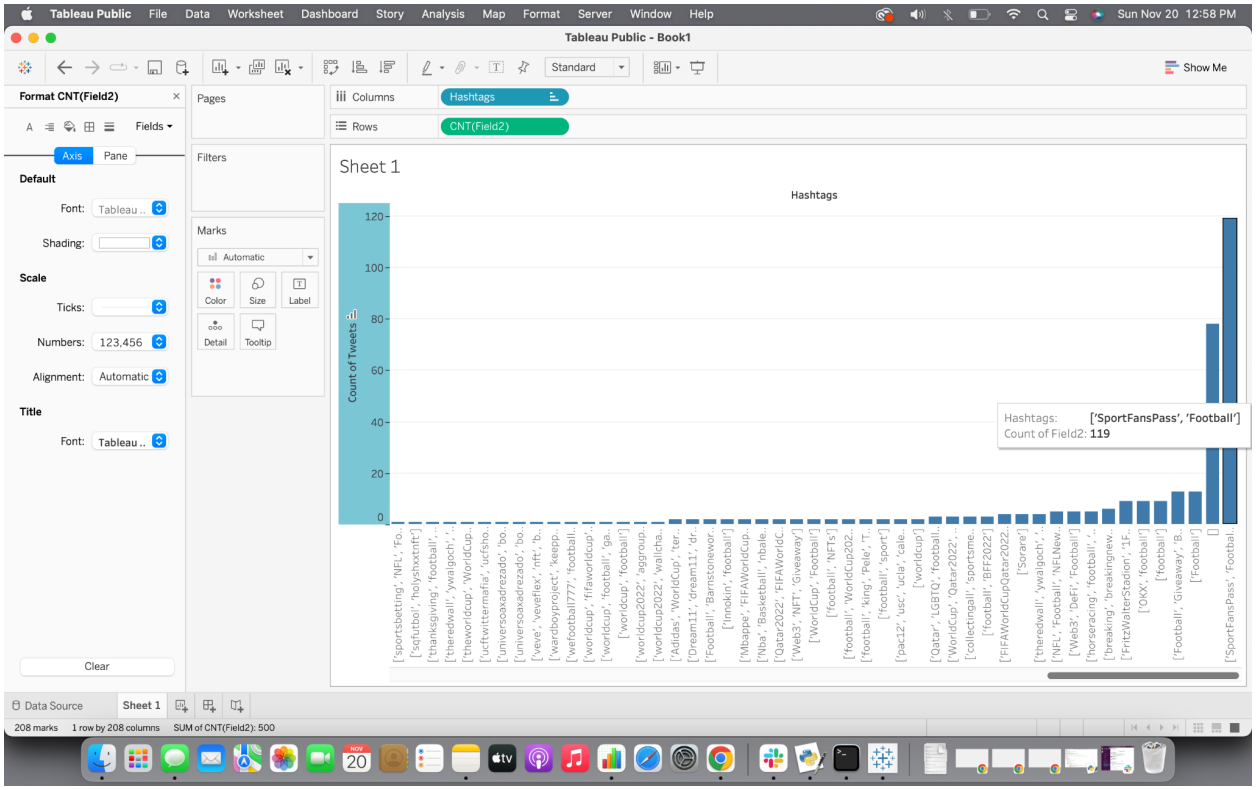
csv_to_store_in_bucket.json

Fields

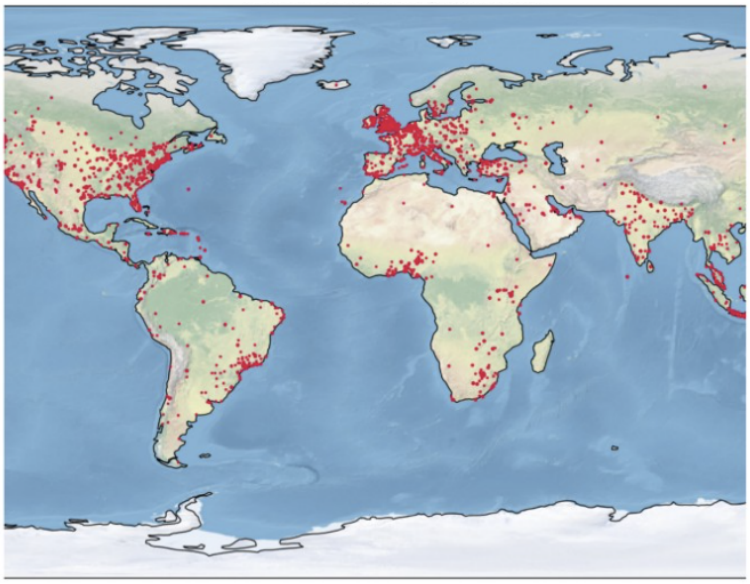
Type	Field Name	Physic...	Rem...
#	Document Index (generated)	csv_to_s...	Docu...
#	Field1	csv_to_s...	FIELD1
#	Field2	csv_to_s...	FIELD2

Abc csv_to_store_in_bucket.json Hashtags	Abc csv_to_store_in_bucket.json Location	# csv_to_store_in_bucket.json Retweetcount	Abc csv_to_store_in_bucket.json Text
[]	Caerdydd | Cardiff	3.00	red wall fest kinds lush pleasu...
['FIFAWorldCupQatar2022', 'Q...	Metaverse	4.00	days kickoff win opener qatar ...
['Colorado', 'Playoffs', 'Footbal...	South Jeffco	0.00	good luck ducks playing quart...
[]	England, United Kingdom	220.00	many congratulations legend ...
['GameFi', 'WorldCupQatar20...		0.00	looking project gamefi excited...
['KBFC', 'Indianfootball', 'footb...		16.00	time far kbfc indianfootball fo...





We see that the count of tweets with respect to difference hashtags contain football with sportsFansPass as this is the data for just a week before Qatar World Cup 2022, and this might be because of the fans looking for the passes and tickets to the game and looking for tickets anywhere and everywhere. Even the total number of retweets were higher for this as well, which might suggest that there might be some giveaways for tickets and passes and people took part in it as they retweeted the post to win the competition.



Above image is the distribution of tweets across the globe and it is clear that most of them are from Europe, as it is the most popular sport in Europe and some of it also from Middle East as the world cup is hosted in Qatar.

Apart from the visualizations and the modeling results, we see we have successfully created a one stop solution for getting insights on the tweets data by creating an end to end pipeline from extraction to results. Apart from that, using technology like PySpark has assured us that even if the volume of the data increases tomorrow or the data processing needs increase in future it will be efficiently able to handle it. Same goes with the AWS services like Athena and Glue, which help in extracting transforming and loading the data a piece of cake and helps a lot in automating the pipeline in just a click and even scheduling. Another advantage now is that we can now query and process just the data that we need rather than loading the whole dataset for processing. Also the big data platforms like AWS gives us an opportunity to host big data applications and there also exists a future scope of hosting different APIs in our VMs which can be accessed through the public IP of the Virtual Machine.

Discussion

When we look at the tweets and count of tweets a week before the world cup we see that hashtags contain football with sportsFansPass, and this might be because of the fans looking for the passes and tickets to the game and looking for tickets anywhere and everywhere. Even the total number of retweets were higher for this as well, which might suggest that there might be some giveaways for tickets and passes and people took part in it as they retweeted the post to win the competition. There is a noticeable concentration of tweets from the East coast of America and western Europe. The Indonesian island of Java has a particularly high concentration of tweets, as well.

So now we have a general idea of the sentiment of the tweets, and how it changes as the world cup came closer. It may also be interesting to delve into the content of the tweets that were made. We will start this by creating a new python file and doing the whole sentiment analysis and report answers to many more new questions but that is out of the scope for this subject's project.

Throughout this course we have learned and worked on the big data management concepts on Google cloud Platform, whether it be virtualization, data lifecycle, storage or modeling, etc. I used all this knowledge that I learned on GCP to apply on AWS. AWS EC2 was used to extract the tweets from twitter and our cloud computing concept was used here. It is similar to Compute Engine in GCP. Data transformation and Loading was performed in AWS Glue with the help of Pyspark, here we covered the concept of parallelization and distributed computing. I also used AWS S3 as my storage service. AWS Athena is used to store the results by querying and connecting AWS S3 to Tableau. After that with Tableau visualization is performed. Amazon Sagemaker is also connected with S3 bucket, as for modeling Sagemaker has very good capabilities. Each and every module that I learned in this course helped me in this project.

One important limitation that I faced while working on AWS was that I had to keep checking what my cost would be when I use each and every service and this even costed me few dollars even though I was being cautious at every step, thus an important lesson as well, I now know where my money went and how should I efficiently use the available data. This will help me in the company as when asked to perform something, I will most probably be vary of the cost associated with it and be most feasible.

Conclusion

We can draw the conclusion that using big data technology is essential for handling ever growing amounts of data. Cluster computing frameworks such as PySpark, highly scalable storage services such as S3, scheduling services which can scale upto the imaginable limit and a highly interactive interpreter language such as Python can help you achieve that. In addition, AWS platform which we used in this proejct has many highly collaborative frameworks and support wide integrations for future scope. We created a big data pipeline from beginning to end for the project. We used a variety of technologies, and there were tradeoffs to be made. I've briefly addressed why I selected each technology or tool in the document above. We learned a lot that we can apply going forward to develop AI-based solutions. In this particular project we used the big data technologies for the analysis of the football tweets and gave a glimpse of what can be achieved with powerful pipelines that can automatically convert the data into something meaningful with a click of a button.

Future Work

In the future, we could put into practice a model that would help us know the trending topic and include that with sentiment analysis to maybe predict if something might go wrong around the world, with some signals that are not captured by human eye, we can probably use AI to our use.

References

- https://medium.com/@jkimera5/creating-real-time-tweets-processing-pipeline-using-aws-ec2-lambda-s3-glue-and-athena-c0004000029d
- https://docs.aws.amazon.com/prescriptive-guidance/latest/patterns/orchestrate-an-etl-pipeline-with-validation-transformation-and-partitioning-using-aws-step-functions.html
- https://medium.com/syntio/streaming-data-from-twitter-to-gcp-7b92c84211a7
- https://medium.com/@tomham000/world-cup-final-2018-twitter-analysis-in-python-2be24e116d59

- https://towardsdatascience.com/how-to-create-a-dataset-with-twitter-and-cloud-computing-fcd82837d313
- https://link.springer.com/article/10.1007/s13278-021-00842-z
- https://aws.amazon.com/datapipeline/details/
- Modules of the course I-535 by Inna Kouper