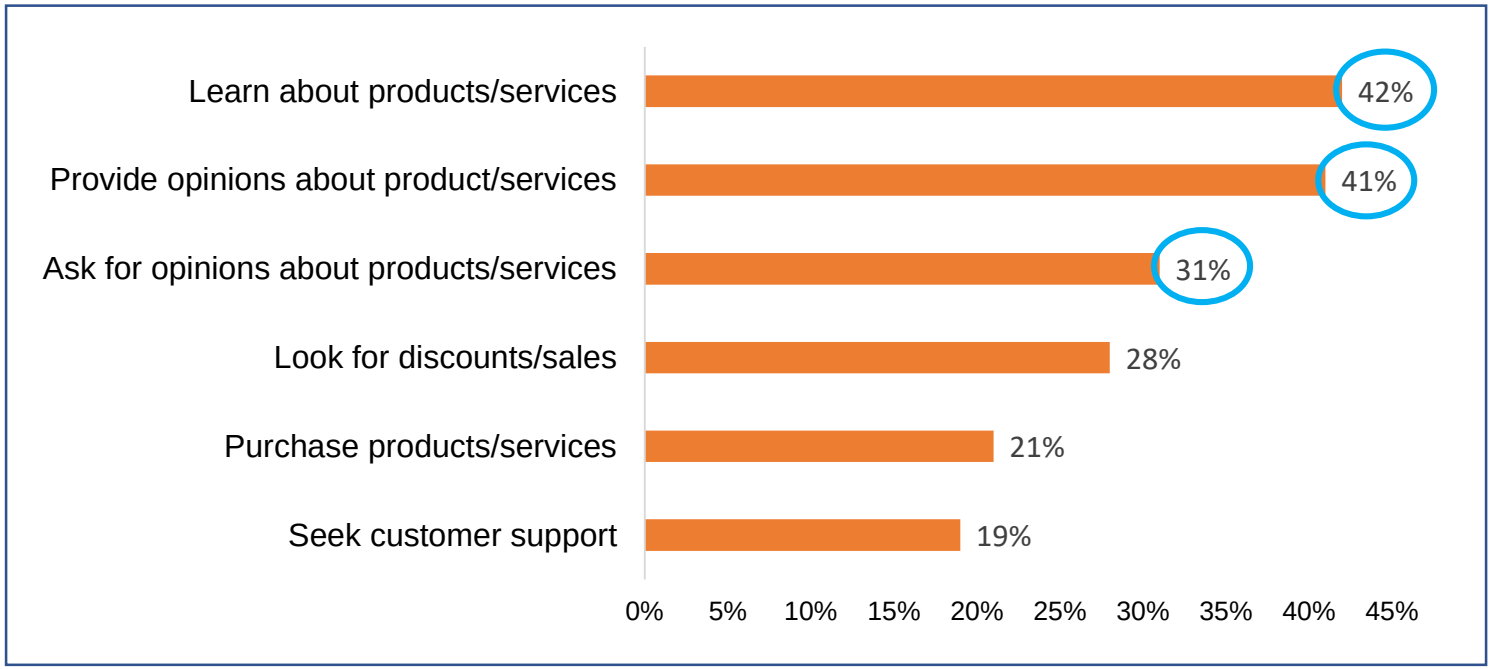
Building a Twitter Big Data Pipeline with Python, Kafka and MongoDB

Manmeet Kumar Chaudhuri

for Businesses/Brands

One of the largest social media platform with c. 335 million Monthly Active Users.

Percentage of Monthly Twitter Users using Twitter to…



65% of the Fortune Global companies have Twitter account.

Source; Edison Research (2014), Burson-Marsteller Study (2014)

The takeaway is clear, the conversation is on and whether or not a brand takes an active part in it, is up to it

for Businesses/Brands

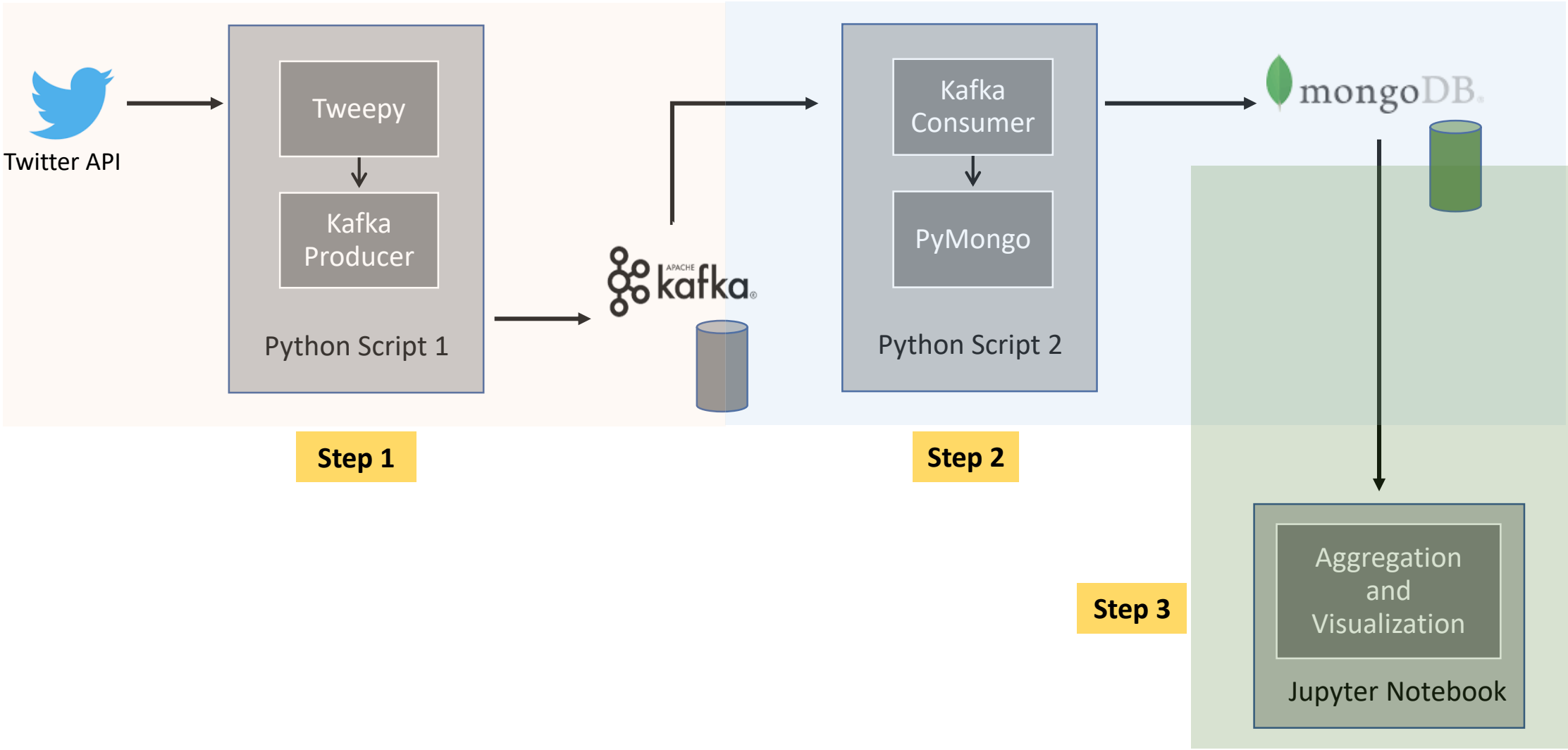
Insights from Twitter Data for a Business/Brand

Total followers of a brand vs. competing brands	Markets/Regions with significant follower base	Trending #hashtags associated with a brand
Real time sentiment of users towards a brand	Preferred languages used by followers for communication	Identify and engage with Influencers for brand endorsement
NLP to further understand customer needs	Followers feedback on products/services	Targeted brand advertisements/increase advertising efficiencies

I will try to analyze a few of these questions from the Twitter Data

Twitter has become an unparalleled source of both structured and unstructured data for Businesses/Brands

Twitter Big Data Pipeline



Data traversing from Twitter to Kafka

Steps used in the process are as follows:

- We will generate the consumer key and the access token for the Twitter API
- Start Zookeeper from the command prompt window:

  zkServer

- We shall now start Kafka by going into the Kafka directory and using the following command:

  bin\windows\kafka-server-start.bat C:\kafka_2.12-2.1.0\kafka_2.12-2.1.0\config\server.properties

- We shall create a topic now. For the purpose of this exercise, we will create the topic “trump” as it is one of the hot and trending topics in the media and Twitter nowadays.

  bin\kafka-topics.bat --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic trump

- We will write a python script [using tweepy library] to get the tweets from the Twitter API

Data traversing from Kafka to MongoDB

Steps used in the process are as follows:

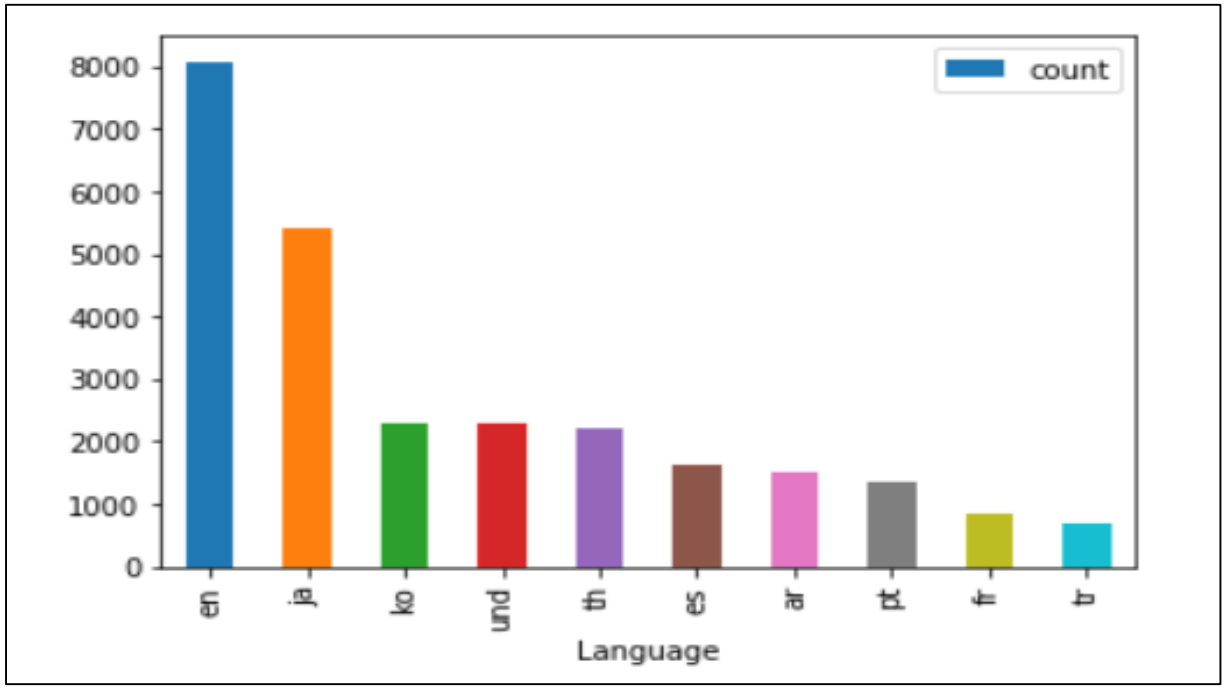
- Zookeeper and Kafka broker should be running
- We shall create a Kafka consumer. The arguments to be used in the Kafka consumer are provided below:
  - The first argument is the topic, *trump* in our case
  - *bootstrap_servers=['localhost:9092']*: same as our producer
  - *auto_offset_reset='earliest'*: one of the most important arguments. It handles where the consumer restarts reading after breaking down or being turned off and can be set either to *earliest* or *latest*. When set to *latest*, the consumer starts reading at the end of the log. When set to *earliest*, the consumer starts reading at the latest committed offset. And that's exactly what we want here
  - *enable_auto_commit=True*: makes sure the consumer commits its read offset every interval
  - *group_id='counters'*: this is the consumer group to which the consumer belongs

- We shall connect to the "trump" collection in our MongoDB database

- We will extract the data from our consumer by looping through it. We will store the extracted data in MongoDB

Data aggregation and visualisation

1. Top 10 languages used by Twitter users for topics related to “trump”



Data aggregation script

```
cursor=tweets.aggregate([
  {"$group":{"_id": "$lang", "count": {"$sum": 1}}},
  {"$sort":{"count": -1}},
  { "$limit": 10 }])
```

Data visualization script using Matplotlib

```
import matplotlib.pyplot as plt
dfr=pd.DataFrame.from_records(cursor)
plot1=dfr.plot(x="_id", y="count", kind="bar")
plot1.xaxis.set_label_text("Language")
```

Data aggregation and visualisation

2. Top 10 hashtags used by Twitter users for topic related to “trump”

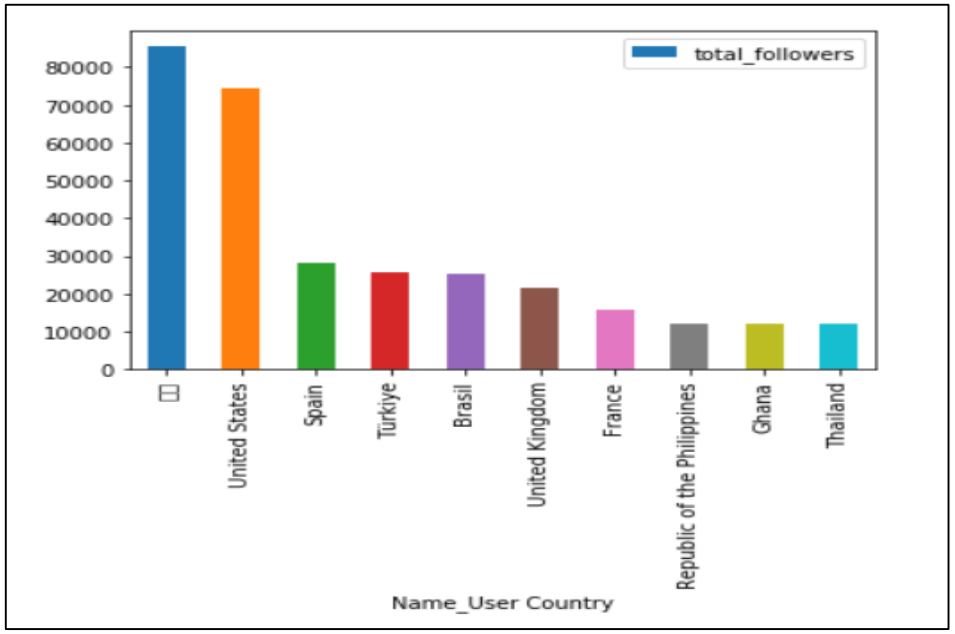
	_id	tagCount
0	방탄소년단	160
1	BTS	152
2	ฮือน่ารัก	125
3	GOT7	65
4	威神V	62
5	WeiShenV	61
6	태형	61
7	TEN	60
8	EXO	59
9	뷔	55

Data aggregation and visualization script

```
cursor1=tweets.aggregate([
{"$unwind": "$entities.hashtags"},
{ "$group": {
"_id": "$entities.hashtags.text", "tagCount": {"$sum": 1}}},
{ "$sort": {"tagCount": -1}},
{ "$limit": 10 }])
dfr2=pd.DataFrame.from_records(cursor1)
dfr2
```

Data aggregation and visualisation

3. Top 10 user countries with the greatest number of followers



Data aggregation script

```
cursor2=tweets.aggregate([
    {"$group":{"_id": "$place.country",
"total_followers":{"$sum":"$user.followers_count"}}},
    {"$sort":{"total_followers": -1}},
    {"$limit": 10}])
```

Data visualization script using Matplotlib

```
import matplotlib.pyplot as plt
dfr=pd.DataFrame.from_records(cursor)
plot1=dfr.plot(x="_id", y="count", kind="bar")
plot1.xaxis.set_label_text("Language")
```

Data aggregation and visualisation

4. Top 10 users with the greatest number of retweets

```
[{'_id': '방탄소년단', 'total_retweets': 11470247},
{'_id': 'スターライトステージ', 'total_retweets': 5144
690}, {'_id': 'c', 'total_retweets': 2976406}, {'_
id': 'Jordan Ireland', 'total_retweets': 1851077},
{'_id': 'David Tra', 'total_retweets': 1530645}, {
'_id': 'juany', 'total_retweets': 1400217}, {'_id'
: 'Maria Fernanda', 'total_retweets': 997301}, {'_
id': 'BTS_official', 'total_retweets': 959946}, {'
_id': 'Aquiles', 'total_retweets': 953754}, {'_id'
: 'BamBam', 'total_retweets': 895634}]
```

Data aggregation script

```
cursor3=tweets.aggregate([
    {"$group":{"_id": "$retweeted_status.user.name",
"total_retweets":{"$sum":"$retweeted_status.retweet_count"}}},
    {"$sort":{"total_retweets": -1}},
    { "$limit": 10 }])
print(list(cursor3))
```

Data aggregation and visualisation

5. Sentiment analysis

- We shall be using TextBlob library. TextBlob is a Python library for processing textual data
- We shall then connect to the Mongo client and access the trump collection
- The tweets are then passed into TextBlob package for sentiment analysis. A polarity score of something in between +1 and -1 is generated for each of the tweets
- A negative score signifies negative sentiment associated with the tweet. A score of 0 means that the users have neutral view of the tweet. A positive score means that that users share a positive opinion of the tweet

RT @imillhiser: This is your friendly reminder that the entire Republican Party opposes democracy and would be happy to see an authoritarian…
0.39166666666666666

RT @dcsportsbog: NFL owners are embarrassed and unhappy about what's been happening in D.C., a league executive told Sally Jenkins, and the…
-0.6

RT @dmightyangel: Whatever happened to USA during Obama's regime will never be credited to Governors of States in America, what is happening…
0.0

RT @dnlbrns: Watching CNN's entire roster get trashed on NYE is quickly becoming one of my favorite traditions https://t.co/79wXnVcSwV
0.31666666666666665

Boho Style Earrings, Copper Bead Earrings, Gifts Under 50, Art Nouveau Style, Romantic Earrings, Boho Chic Earrings… https://t.co/4prFtK59Uv
0.0

Difficulties/challenges faced in the project

A few difficulties/challenges faced in the project are as follows:

- Integrating python script with Twitter API and Kafka, Kafka and MongoDB
- Usage of “matplotlib” library in python
- Aggregation functions in MongoDB


AZURE
BYOD?
CLOUD
SECURITY
BIG DATA
PATTERNS
DEVELOPERS
SOCIAL
APPS
PLATFORM
WEB
MAGIC
DEVICES
MEMORY
PEOPLE

Thank you!