Twitter Pipeline

- ## Overview
  We want to get data from Twitter considering a topic and making some transformations from the data and gaining insights with spark then we will store it in our HDFS then with hive we will create DataWarehouse Where we will have dimensions and facts table.

- ## Twitter Listener
  Our first step in the process is making a script that connects to Twitter API and retrieves recent tweets that match a given search query
  And our topic is “Cleopatra” Then we retrieved the data needed for our work or what we were interested in with a Get request after that we converted it to a JSON object and then send it to the socket for the next script.

- ## Twitter_Stream
  It’s the second step in our pipeline it’s pyspark script reads data from the socket stream after that we defined the field we are interested in and made a schema for it using Struct Type the data we Retrieved are
  Tweet: id, text,in_reply_to_user_id,created_at,public_metrics, source.
  User: username,user_metrics.
  In explaining they are the tweet id and it’s text,we also retrieved in_reply to user to see id this tweet is reply or new tweet,created at is the timestamp and source is from where is tweet is made(mobile,pc,...), and the tweet metrics which have (retweet count,replycount,likecount,....).
  For user, i choose to retrieve the username not the name because that is the unique one and we retrieved the user metrics (followers, following).
  After that, we converted the JSON to Dataframe using our Schema .
  After that we made some cleaning and transformation, we separated the public and user metrics into their fields,also for the in_reply to i converted the column to boolean with true and false,for the source we made it to be unknown in case there’s no source available, we extracted hour,day,month,year from the timestamp of the tweet.
  Lastly, we stored the data frame in parquet file and created a hive table on top it to be the landing table(staging table)

here it was stated that we create the table on top of it but if we made the parquet append then the landing table will have too much data which I think isn't correct it should be used to store data till insert into it's tables also if we made overwrite we will have no copy of the original data so i choose to store parque with append as backup and another one as overwrite to be for the landing table and choose it to be parquet instead of just inserting it so we can move it if need in any other thing

- ## Hive Dimention tables

  The third step is creating dimension tables and inserting the data into them i created 2 dimension tables tweet_raw that have all the data about tweets and user_raw which have all the data about users.
  While inserting the data i made sure to have no duplicates so when inserting the data i made sure to check if this tweet is in the tweet_raw before or not,in user_raw it's the same
  the user_raw should be updated but must decide how often because if it's not updated the users may have much more following and if we update it when there's a change it will be updated too much because every tweet he may gain followers and so as we now in datawarehouse we don't update to keep record of change so the old record must be set to high date before inserting the new one

- ## Fact tables

  It's python script that create fact tables from the dimension tables I choose that the fact grain is day to have a meaning because anyone that will want to use it rarely will he follow it hour by hour but it can be done.
  I made sure that there's no duplicates when appending .
  First the users_fact table is used to get the coverage like if we want to get how much did this topic get coverage how many did it reach for example if 5 users tweeted about it each one have 1m follower so we can assume that the coverage of it it 5m views.
  So imade sure when geeting the users in fact i choose not all users but the one that maded tweets today and if he made more than a tweet he will only be counted once and if ran the the script and it hade the same tweets it won't insert them again

- Pipeline
  The last script is the script we collect in all the scripts we want to run it could be made as bash script or python script

Notes:-

- I know the script we wanted it to run every 10min or so but it almost didn't bring any data so I made it daily it could be turned to each 10 min and we can change the fact grain to an hour then make a fact over it for daily
- I choose Cleopatra topic because it's controversial as i planned to make a sentiment analysis and divide them to posit and negative data to be separate fact tables for each and make another fact to measure them put only for influencers(defined to have a certain number of followers)and another one for famous tweets that have many retweets ,likes,replies. But unfortunately, i was swamped and couldn't apply that