Bigdata Assignment 8.3

Counting popular hashtags using Spark sql

Below is the sample tweet

{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":"FilmFan","truncated":false,"lang":"en","in_reply_to_status_id_str":null,"id":689085590822891521,"in_reply_to_user_id_str":"6048122","timestamp_ms":"1453125782100","in_reply_to_status_id":null,"created_at":"Mon Jan 18 14:03:02 +0000 2016","favorite_count":0,"place":null,"coordinates":null,"text":"@filmfan hey its time for you guys follow @acadgild To #AchieveMore and participate in contest Win Rs.500 worth vouchers","contributors":null,"geo":null,"entities":{"symbols":[],"urls":[],"hashtags":[{"text":"AchieveMore","indices":[56,68]}],"user_mentions":[{"id":6048122,"name":"Tanya","indices":[0,8],"screen_name":"FilmFan","id_str":"6048122"},{"id":2649945906,"name":"ACADGILD","indices":[42,51],"screen_name":"acadgild","id_str":"2649945906"}]},"is_quote_status":false,"source":"<a href=\"https://about.twitter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck<\/a>","favorited":false,"in_reply_to_user_id":6048122,"retweet_count":0,"id_str":"689085590822891521","user":{"location":"India ","default_profile":false,"profile_background_tile":false,"statuses_count":86548,"lang":"en","profile_link_color":"94D487","profile_banner_url":"https://pbs.twimg.com/profile_banners/197865769/1436198000","id":197865769,"following":null,"protected":false,"favourites_count":1002,"profile_text_color":"000000","verified":false,"description":"Proud Indian, Digital Marketing Consultant,Traveler, Foodie, Adventurer, Data Architect, Movie Lover, Namo Fan","contributors_enabled":false,"profile_sidebar_border_color":"000000","name":"Bahubali","profile_background_color":"000000","created_at":"Sat Oct 02 17:41:02 +0000 2010","default_profile_image":false,"followers_count":4467,"profile_image_url_https":"https://pbs.twimg.com/profile_images/664486535040000000/GOjDUiuK_normal.jpg","geo_enabled":true,"profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png","follow_request_sent":null,"url":null,"utc_offset":19800,"time_zone":"Chennai","notifications":null,"profile_use_background_image":false,"friends_count":810,"profile_sidebar_fill_color":"000000","screen_name":"Ashok_Uppuluri","id_str":"197865769","profile_image_url":"http://pbs.twimg.com/profile_images/664486535040000000/GOjDUiuK_normal.jpg","listed_count":50,"is_translator":false}}

Solution -

- tweets are read as a json file and it is registered as temporary table 'tweets'.

```
val tweets = spark.read.json("file:///home/acadgild/Downloads/tweets").registerTempTable("tweets")
```

- Id , tweets's text is selected from the tweets table and then it is made into 'hashtags' table.

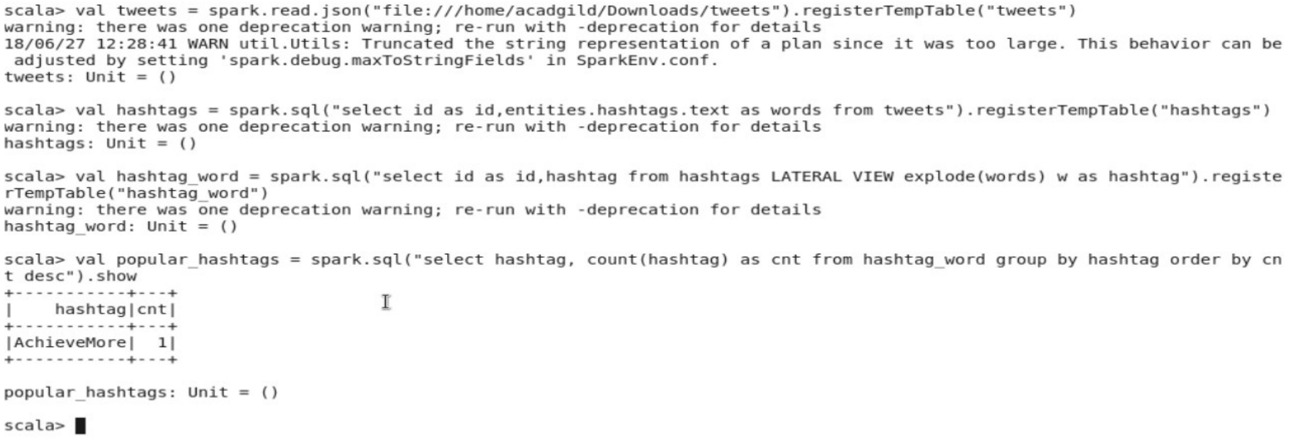
  val hashtags = spark.sql("select id as id,entities.hashtags.text as words from tweets").registerTempTable("hashtags")

- id , hastags are selected and each for each new hasttag a row is created. This is stored as a temporary table 'hashtag_word'.

  val hashtag_word = spark.sql("select id as id,hashtag from hashtags LATERAL VIEW explode(words) w as hashtag").registerTempTable("hashtag_word")

- For each hashtag , it is grouped to find its count and its sorted in descending order.

val popular_hashtags = spark.sql("select hashtag, count(hashtag) as cnt from hashtag_word group by hashtag order by cnt desc").show



```
scala> val tweets = spark.read.json("file:///home/acadgild/Downloads/tweets").registerTempTable("tweets")
warning: there was one deprecation warning; re-run with -deprecation for details
18/06/27 12:28:41 WARN util.Utils: Truncated the string representation of a plan since it was too large. This behavior can be
 adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
tweets: Unit = ()

scala> val hashtags = spark.sql("select id as id,entities.hashtags.text as words from tweets").registerTempTable("hashtags")
warning: there was one deprecation warning; re-run with -deprecation for details
hashtags: Unit = ()

scala> val hashtag_word = spark.sql("select id as id,hashtag from hashtags LATERAL VIEW explode(words) w as hashtag").registe
rTempTable("hashtag_word")
warning: there was one deprecation warning; re-run with -deprecation for details
hashtag_word: Unit = ()

scala> val popular_hashtags = spark.sql("select hashtag, count(hashtag) as cnt from hashtag_word group by hashtag order by cn
t desc").show
+-----------+---+
|    hashtag|cnt|
+-----------+---+
|AchieveMore|  1|
+-----------+---+

popular_hashtags: Unit = ()

scala> 
```

Output -

As there was only 1 tweets , so 1 hashtag is shown. But it will work for a larger sample also.