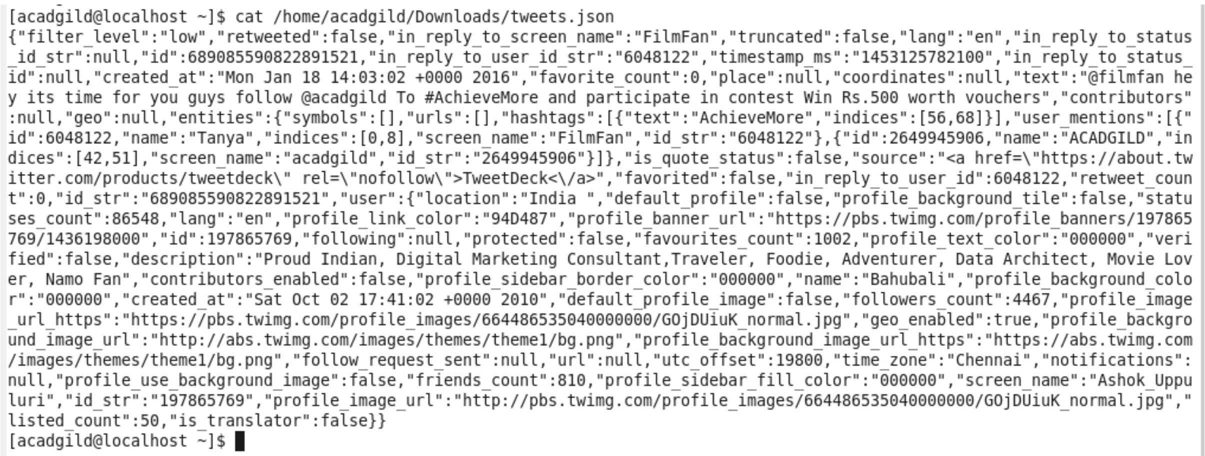
Problem Statement

Implement the below blog at your end and send the complete documentation.
https://drive.google.com/file/d/0B_Qjau8wv1KobUlaOEtfNEtQNkU/view?usp=sharing

(Counting popular hashtags using spark.odt)

Dataset:

Snapshot of tweets collected as dataset

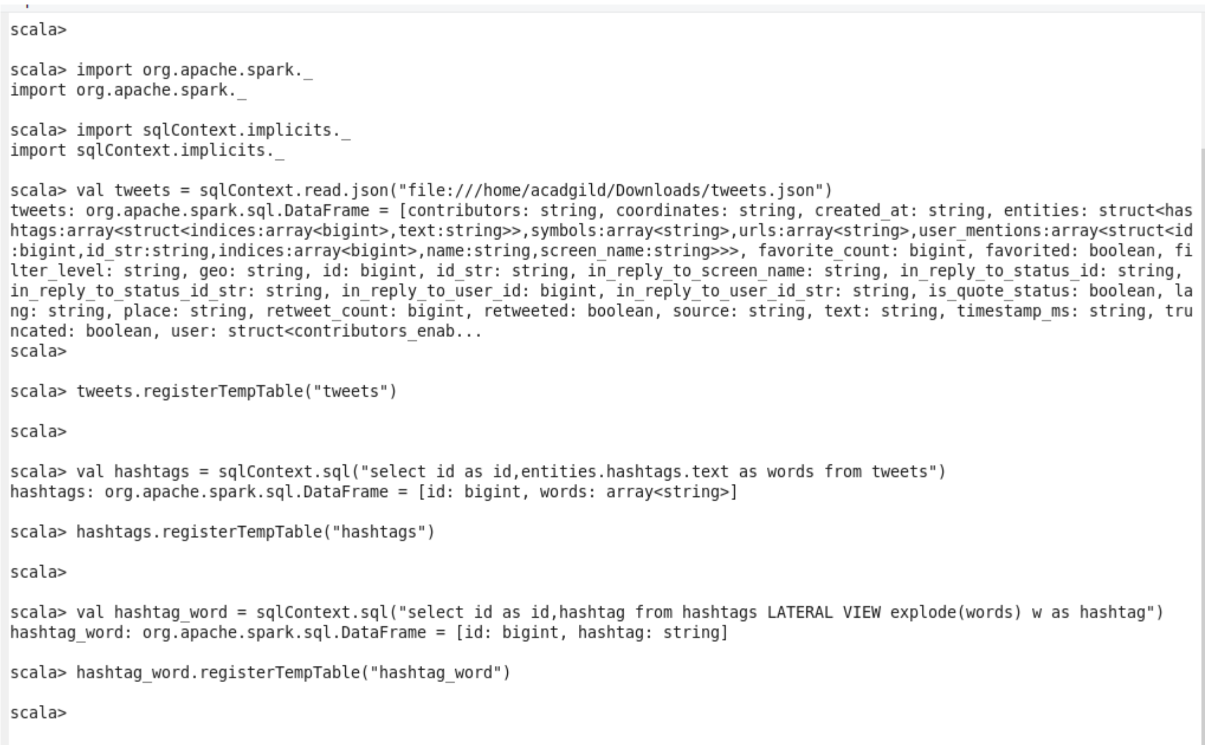


```
[acadgild@localhost ~]$ cat /home/acadgild/Downloads/tweets.json
{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":"FilmFan","truncated":false,"lang":"en","in_reply_to_status
_id_str":null,"id":689085590822891521,"in_reply_to_user_id_str":"6048122","timestamp_ms":"1453125782100","in_reply_to_status_
id":null,"created_at":"Mon Jan 18 14:03:02 +0000 2016","favorite_count":0,"place":null,"coordinates":null,"text":"@filmfan he
y 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","in
dices":[42,51],"screen_name":"acadgild","id_str":"2649945906"}]},"is_quote_status":false,"source":"<a href=\"https://about.tw
itter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck<\/a>","favorited":false,"in_reply_to_user_id":6048122,"retweet_coun
t":0,"id_str":"689085590822891521","user":{"location":"India ","default_profile":false,"profile_background_tile":false,"statu
ses_count":86548,"lang":"en","profile_link_color":"94D487","profile_banner_url":"https://pbs.twimg.com/profile_banners/197865
769/1436198000","id":197865769,"following":null,"protected":false,"favourites_count":1002,"profile_text_color":"000000","veri
fied":false,"description":"Proud Indian, Digital Marketing Consultant,Traveler, Foodie, Adventurer, Data Architect, Movie Lov
er, Namo Fan","contributors_enabled":false,"profile_sidebar_border_color":"000000","name":"Bahubali","profile_background_colo
r":"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_backgro
und_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_Uppu
luri","id_str":"197865769","profile_image_url":"http://pbs.twimg.com/profile_images/664486535040000000/GOjDUiuK_normal.jpg","
listed_count":50,"is_translator":false}}
[acadgild@localhost ~]$ █
```

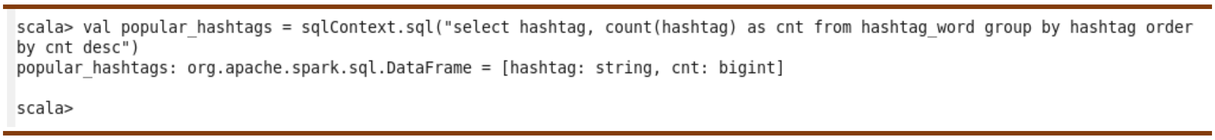
Solution:

- Import Spark sql packages
- Begin by reading the data file as a json file from the local FS using the SQL context object ***sqlContext***.
- This data is then used to create a temporary table tweets. This is variable tweets.
- From the table **tweets** created above, Select the **id** and the **text** (words) (from the hashtags element in entities column)
- This data is then used to create a temporary table hashtags. This is variable hashtags.
- From the table **hashtags** created above, Select the **id** and a column called **hashtag** that is created by using LATERAL VIEW explode function with the column **words**:
- This function takes the column **words** (which has multiple elements) as argument and separates every element; creating a new row for each.
- Ex: (1, “Hi Hello How”) becomes (1, “Hi”), (1, “Hello”), (1, “How”)
- This data is then used to create a temporary table hashtag_word. This is variable hashtag_word.
- Finally to get the count for popular hashtags,
- Group the data by the hashtag. This will give a group of data corresponding to every hashtag.
- For every group, Select the name of the **hashtag** and the count of the hashtag in the group (cnt)
- The data is lastly ordered by hashtag with the highest count first (DESC).

Code:



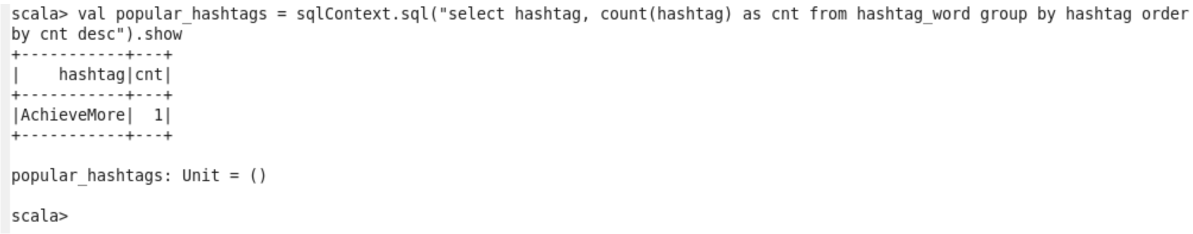
```
scala>

scala> import org.apache.spark._
import org.apache.spark._

scala> import sqlContext.implicits._
import sqlContext.implicits._

scala> val tweets = sqlContext.read.json("file:///home/acadgild/Downloads/tweets.json")
tweets: org.apache.spark.sql.DataFrame = [contributors: string, coordinates: string, created_at: string, entities: struct<has
htags:array<struct<indices:array<bigint>,text:string>>,symbols:array<string>,urls:array<string>,user_mentions:array<struct<id
:bigint,id_str:string,indices:array<bigint>,name:string,screen_name:string>>>, favorite_count: bigint, favorited: boolean, fi
lter_level: string, geo: string, id: bigint, id_str: string, in_reply_to_screen_name: string, in_reply_to_status_id: string,
in_reply_to_status_id_str: string, in_reply_to_user_id: bigint, in_reply_to_user_id_str: string, is_quote_status: boolean, la
ng: string, place: string, retweet_count: bigint, retweeted: boolean, source: string, text: string, timestamp_ms: string, tru
ncated: boolean, user: struct<contributors_enab...
scala>

scala> tweets.registerTempTable("tweets")

scala>

scala> val hashtags = sqlContext.sql("select id as id,entities.hashtags.text as words from tweets")
hashtags: org.apache.spark.sql.DataFrame = [id: bigint, words: array<string>]

scala> hashtags.registerTempTable("hashtags")

scala>

scala> val hashtag_word = sqlContext.sql("select id as id,hashtag from hashtags LATERAL VIEW explode(words) w as hashtag")
hashtag_word: org.apache.spark.sql.DataFrame = [id: bigint, hashtag: string]

scala> hashtag_word.registerTempTable("hashtag_word")

scala>
```



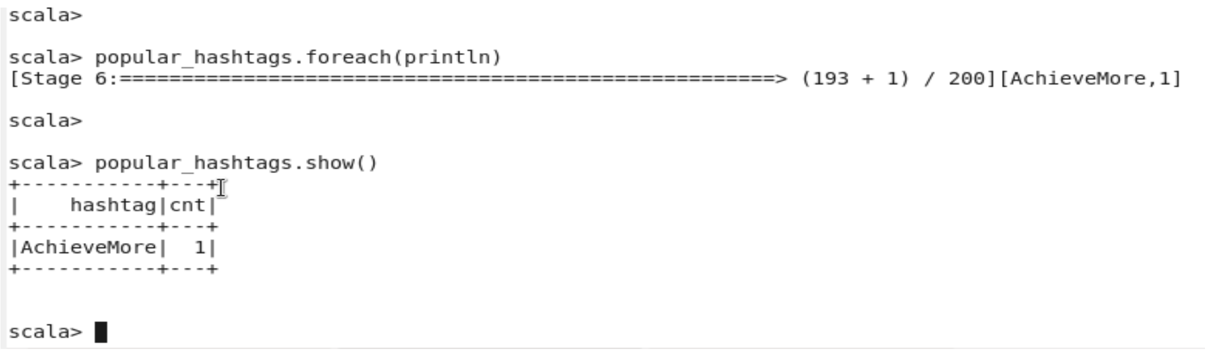
```
scala> val popular_hashtags = sqlContext.sql("select hashtag, count(hashtag) as cnt from hashtag_word group by hashtag order
by cnt desc")
popular_hashtags: org.apache.spark.sql.DataFrame = [hashtag: string, cnt: bigint]

scala>
```

Output:

Output can be shown by using show or foreach method



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

popular_hashtags: Unit = ()

scala>
```



```
scala>

scala> popular_hashtags.foreach(println)
[Stage 6:========================================================> (193 + 1) / 200][AchieveMore,1]

scala>

scala> popular_hashtags.show()
+-----------+---+
|    hashtag|cnt|
+-----------+---+
|AchieveMore|  1|
+-----------+---+


scala> █
```