ASSIGNMENT 21.1

Input Dataset:

Tweets – json file

tweets


```
{
  "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",
```

```
          "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/GOiDUiuK_normal.jpg",
    "listed_count": 50,
    "is_translator": false
}
```


Here is the Spark code snippet to count popular hash tags using Spark SQL:

```
// import required Spark packages
import org.apache.spark.sql.SparkSession

object Assignment21_1 {

  def main(args: Array[String]): Unit = {

// create a SparkSession object that can be used to create various contexts of Spark such as sqlContext
    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()

// read json file and convert the data into a temporary table – 'tweets'

    val tweets = spark.read.json("file:///E:/Acadgild/Session 21/assignment 21.1/ tweets") .registerTempTable("tweets")

// SQL query to fetch tweet id and hashtags present in tweets and convert it into a table – 'hashtags'

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

// SQL query to separate only the hashtags and store them into 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")

// SQL query to get hashtags and their count. Print result to the console

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

}

}

Assignment21_1.scala

```scala
package org.spark_samples

import org.apache.spark.sql.SparkSession

object Assignment21_1 {

 def main(args: Array[String]) {

  val spark = SparkSession
    .builder()
    .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
    .master("local[*]")
    .getOrCreate()

  val tweets = spark.read.json("file:///E:\\Acadgild\\Session 21\\assignment 21.1\\tweets").registerTempTable("tweets")

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

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

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

}
```

Output:

```
+-----------+---+
|    hashtag|cnt|
+-----------+---+
|AchieveMore|  1|
+-----------+---+
```

To conclude, there is only one hashtag in the given json file of tweets, **'AchieveMore'** which has occurred once.