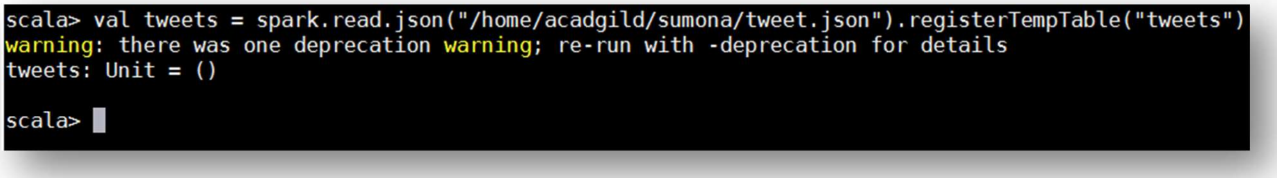
Assignment 21.1

We have created a **json** file names tweet.json and updated in the local file system **/home/acadgild/Sumona**

Following are few commands to get the popular hashtags used in twitter

1. First we will read the JSON file stored in the local file system and create a temporary table **tweets**

val tweets = spark.read.json("/home/acadgild/sumona/tweet.json").registerTempTable("tweets")



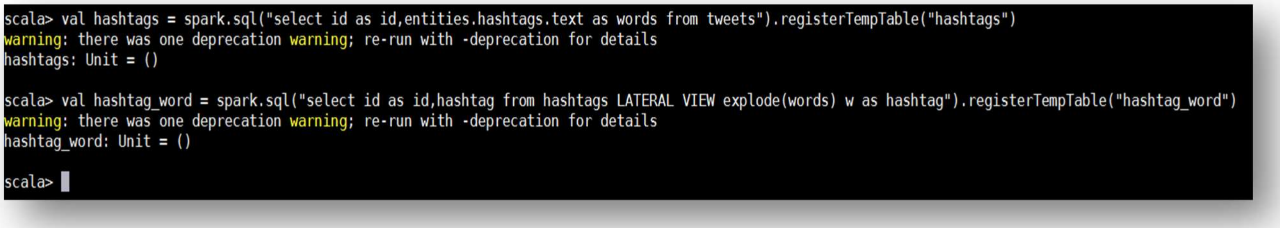
```
scala> val tweets = spark.read.json("/home/acadgild/sumona/tweet.json").registerTempTable("tweets")
warning: there was one deprecation warning; re-run with -deprecation for details
tweets: Unit = ()

scala> 
```

2. Now, from the above temporary table we will select the ID’s. hashtag and create another temporary table **hashtags**

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")



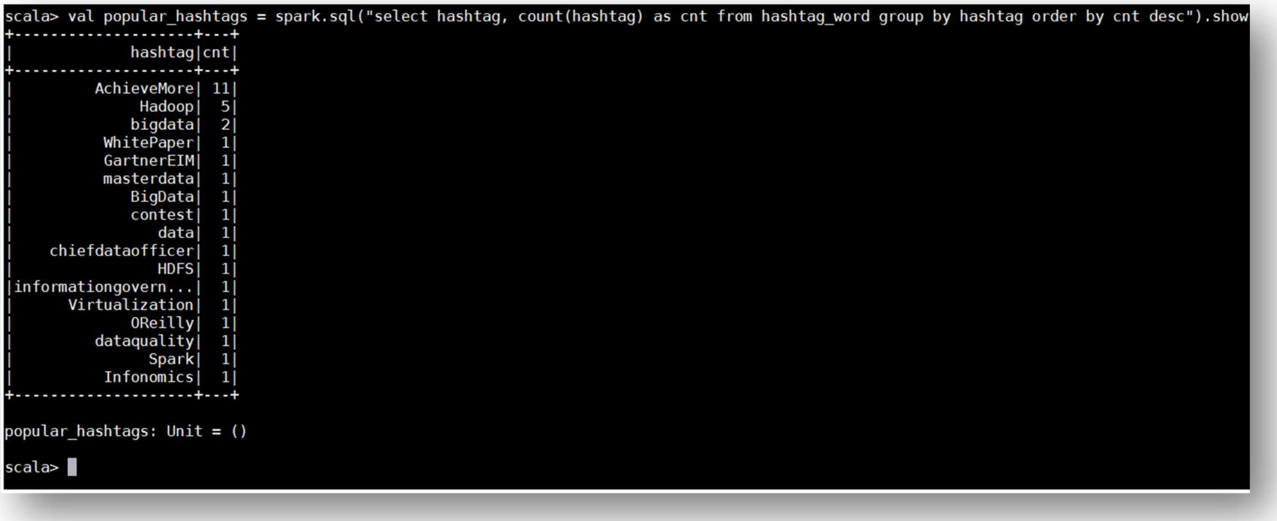
```
scala> val hashtags = spark.sql("select id as id,entities.hashtags.text as words from tweets").registerTempTable("hashtags")
warning: there was one deprecation warning; re-run with -deprecation for details
hashtags: Unit = ()

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

scala> 
```

3. Finally, we will get the popular hashtags used in twitter and its count with the following command:

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



```
scala> val popular_hashtags = spark.sql("select hashtag, count(hashtag) as cnt from hashtag_word group by hashtag order by cnt desc").show
+--------------------+---+
|             hashtag|cnt|
+--------------------+---+
|         AchieveMore| 11|
|              Hadoop|  5|
|             bigdata|  2|
|          WhitePaper|  1|
|          GartnerEIM|  1|
|          masterdata|  1|
|             BigData|  1|
|             contest|  1|
|                data|  1|
|     chiefdataofficer|  1|
|                HDFS|  1|
|informationgovern...|  1|
|      Virtualization|  1|
|             OReilly|  1|
|         dataquality|  1|
|               Spark|  1|
|          Infonomics|  1|
+--------------------+---+

popular_hashtags: Unit = ()

scala> 
```