Assignment 21.1:

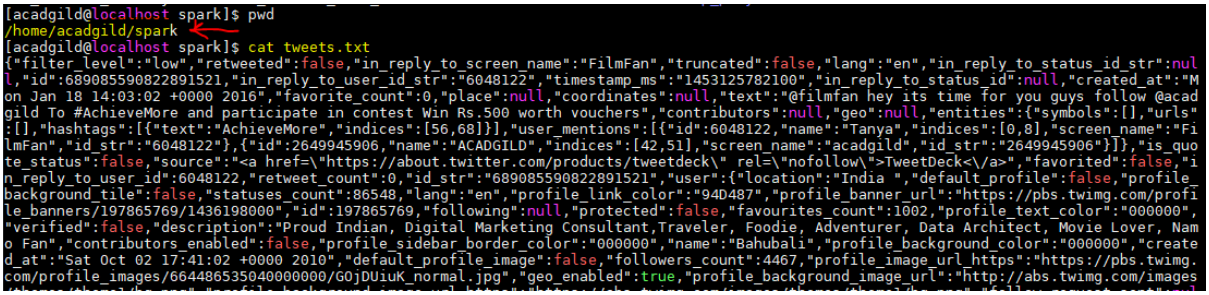
Problem Statement:

Implement the below blog at your end and send the complete documentation.

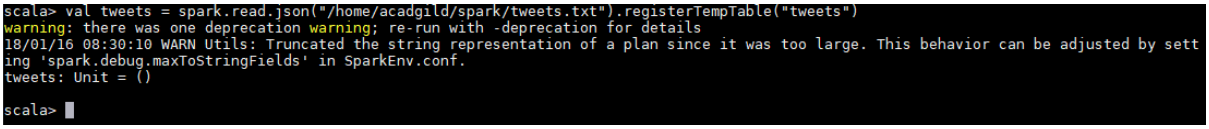
https://drive.google.com/file/d/0B_Qjau8wv1KobUlaOEtfNEtQNkU/view?usp=sharing

Solution:

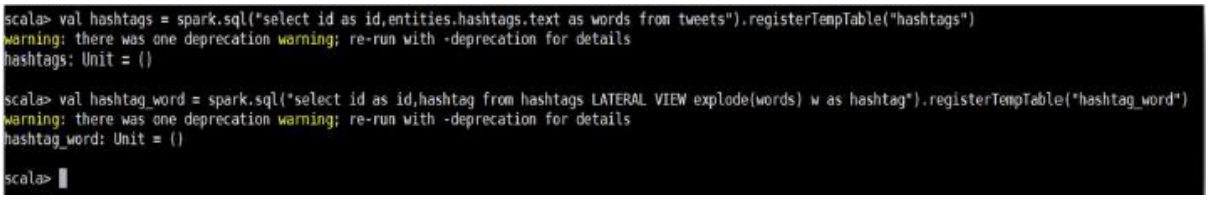
1. Data set is places at below location:



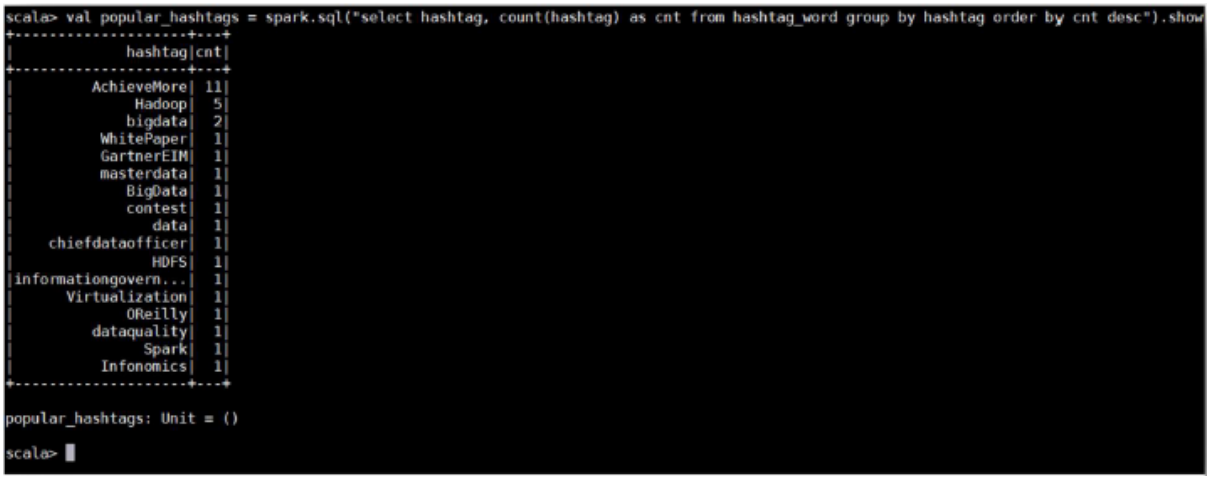
2. First we will read the JSON file stored in the local file system and create a temporary table tweets
   val tweets =
   spark.read.json("/home/acadgild/spark/tweets.txt").registerTempTable("tweets")



3. 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")



4. 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>