**Session 21: MLIB AND GRAPHX**

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](https://drive.google.com/file/d/0B_Qjau8wv1KobUlaOEtfNEtQNkU/view?usp=sharing%20)

Counting popular hashtags using Spark sql

**Counting popular hashtags using Spark sql**

Below is the sample tweet, we uploaded the below tweets as a .json file to the local file system,

{"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}}

The file is uploaded as Counting\_popular\_hashtags\_using\_spark.json into the below location,

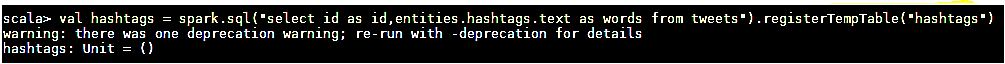
***/home/acadgild/hadoop/Counting\_popular\_hashtags\_using\_spark.json***

**Codes**

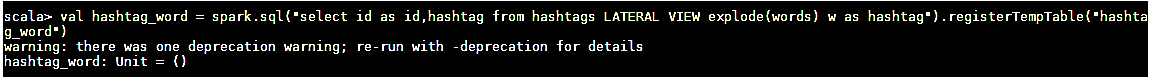
***val tweets = spark.read.json("file:///home/acadgild/hadoop/Counting\_popular\_hashtags\_using\_spark.json").registerTempTable("tweets")***

******

***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**

