1. **Getting hashtags and urls (url, expanded\_url,display\_url) using hive**

Loaded twitter json file using hive to twitter\_table and retrieved hashtags,urls (url,expanded\_url,display\_url ) , please find code details

* create tabke twitter\_table (json string);
* load data local inpath ‘/home/raji/Downloads/tweets.json’ INTO TABLE twitter\_table;
* LOAD DATA LOCAL INPATH  '/tmp/simple.json' INTO TABLE json\_table;
* INSERT OVERWRITE DIRECTORY '/test/proj1\_hashtags'

select explode(split(substr(get\_json\_object(twitter\_table.json, '$.entities.hashtags.text'),2,length(get\_json\_object(twitter\_table.json, '$.entities.hashtags.text')) - 2) , ',') ) as hashtags from twitter\_table

where get\_json\_object(twitter\_table.json, '$.entities.hashtags.text') is not null ;

* INSERT OVERWRITE DIRECTORY '/test/proj1\_url'

select get\_json\_object(twitter\_table.json, '$.entities.urls.url')

from twitter\_table

where get\_json\_object(twitter\_table.json, '$.entities.urls.url') is not null;

* INSERT OVERWRITE DIRECTORY '/test/proj1\_expanded\_url'

select get\_json\_object(twitter\_table.json, '$.entities.urls.expanded\_url')

from twitter\_table

where get\_json\_object(twitter\_table.json, '$.entities.urls.expanded\_url') is not null;

* INSERT OVERWRITE DIRECTORY '/test/proj1\_display\_url'

select get\_json\_object(twitter\_table.json, '$.entities.urls.display\_url')

from twitter\_table

where get\_json\_object(twitter\_table.json, '$.entities.urls.display\_url') is not null;

2. Merged all the files to /test/proj1\_output/hashtags\_urls

3. Word count program Apache Hadoop

hadoop jar hadoop/hadoop-2.8.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.1.jar wordcount /test/proj1\_output/hashtags\_urls /test/projwc

4. Word count program Apache Spark

scala > var hturl\_count=sc.textFile("/home/raji/Downloads/PROJ1/flying\_squarels\_hashtags\_urls.txt")

scala > val hucounts=hturl\_count.flatMap(\_.split(" ")).map(word => (word, 1)).reduceByKey(\_+\_)

hucounts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[13] at reduceByKey at <console>:26

scala> hucounts.collect()

res2: Array[(String, Int)] = Array((["https://t.co/hADR8cFOBi"],1), (["instagram.com/p/BZpQXXBnaMq/"],1), (["https://www.instagram.com/p/BZpaBVPAG8XYxyCBVtYHboumPquSggQjj3DxG80/"],1), (["instagram.com/p/BZpRq24nU\_W/"],1), (["https://t.co/zl7ULo9ZWi"],1), (["http://untp.beer/s/c509442812"],1), (["instagram.com/p/BZpa6IFjDyo/"],1), (["Twitch.Tv/SwaysTwitch","twitter.com/i/web/status/9…"],1), (["https://t.co/qBEcutdzkw"],1), (["https://t.co/DZdEpRQxUG"],1), (["https://www.instagram.com/p/BZpdbruDKSi/"],1), (["goo.gl/fTHjrC","tbtalks.com"],1), (["instagram.com/p/BZpYzmSn1oE\_…"],1), ("ALDUBKapitHigpit",1), (["https://t.co/vFyjZDif8p"],1), (["instagram.com/p/BZpbYYEnhiP/"],1), ("Syria",10), (["https://t.co/zj34WYgley"],1), ("storms",1), (["https://t.co/YhIgAmWelo"],1), (["https://t.co/60tWz75...

scala> hucounts.saveAsTextFile("/test/sparkwc.txt")

**OUTPUT/HADOOP LOG FILES**

https://github.com/PranoopMutha/CS5540\_PB\_FlyingSquirrels\_TwitterProject

**REFERNECES**

http://thornydev.blogspot.com/2013/07/querying-json-records-via-hive.html