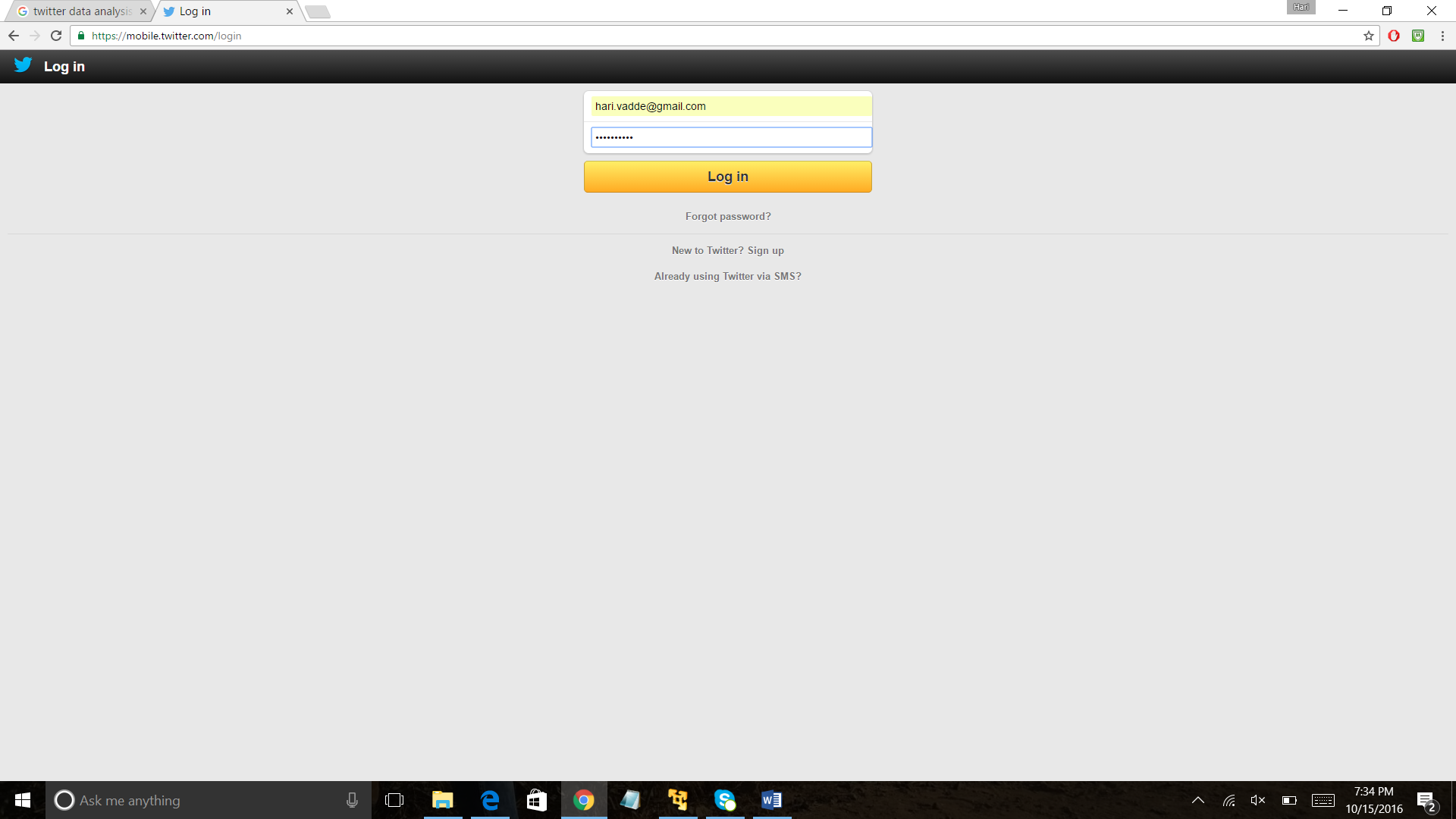
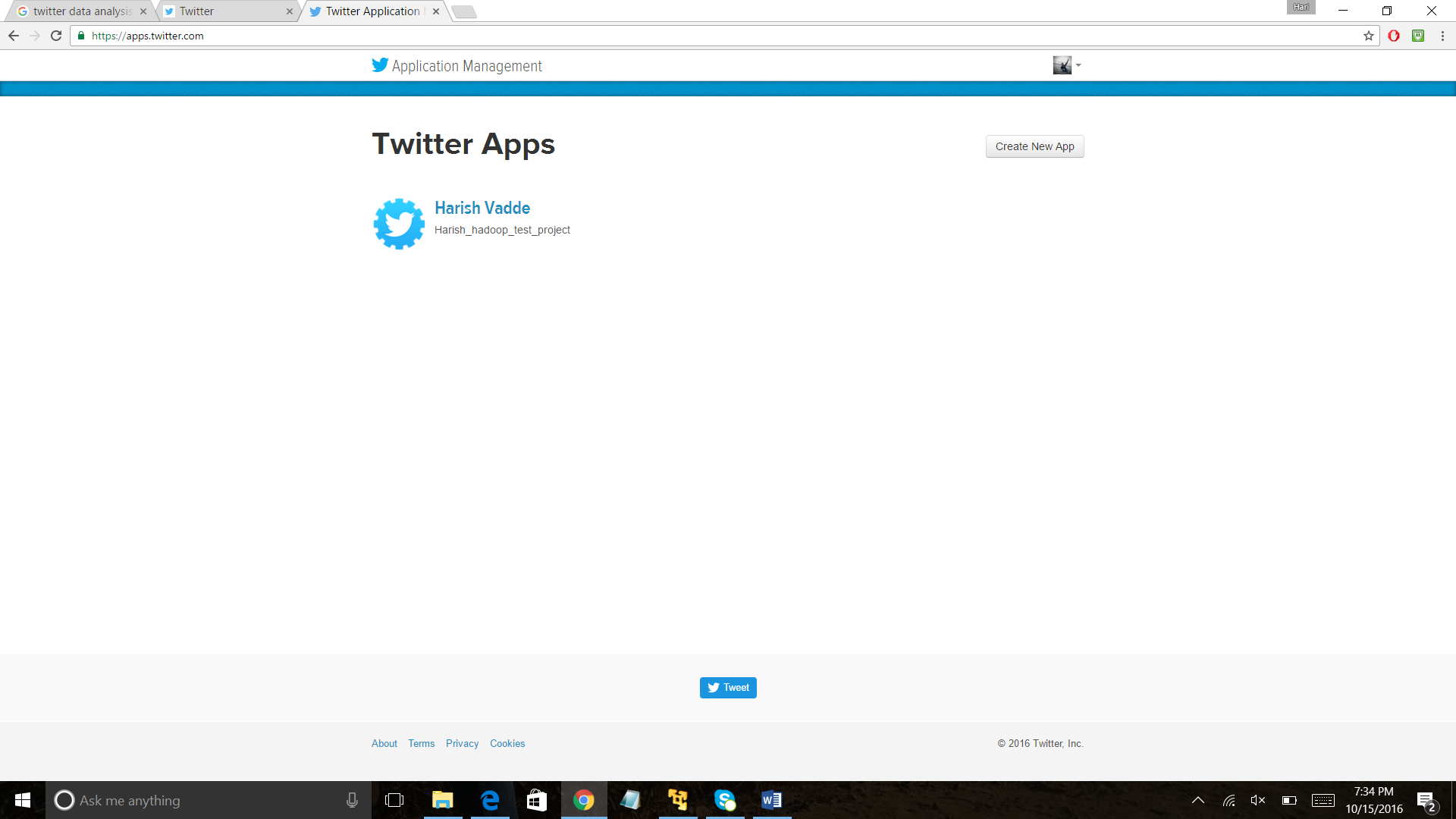
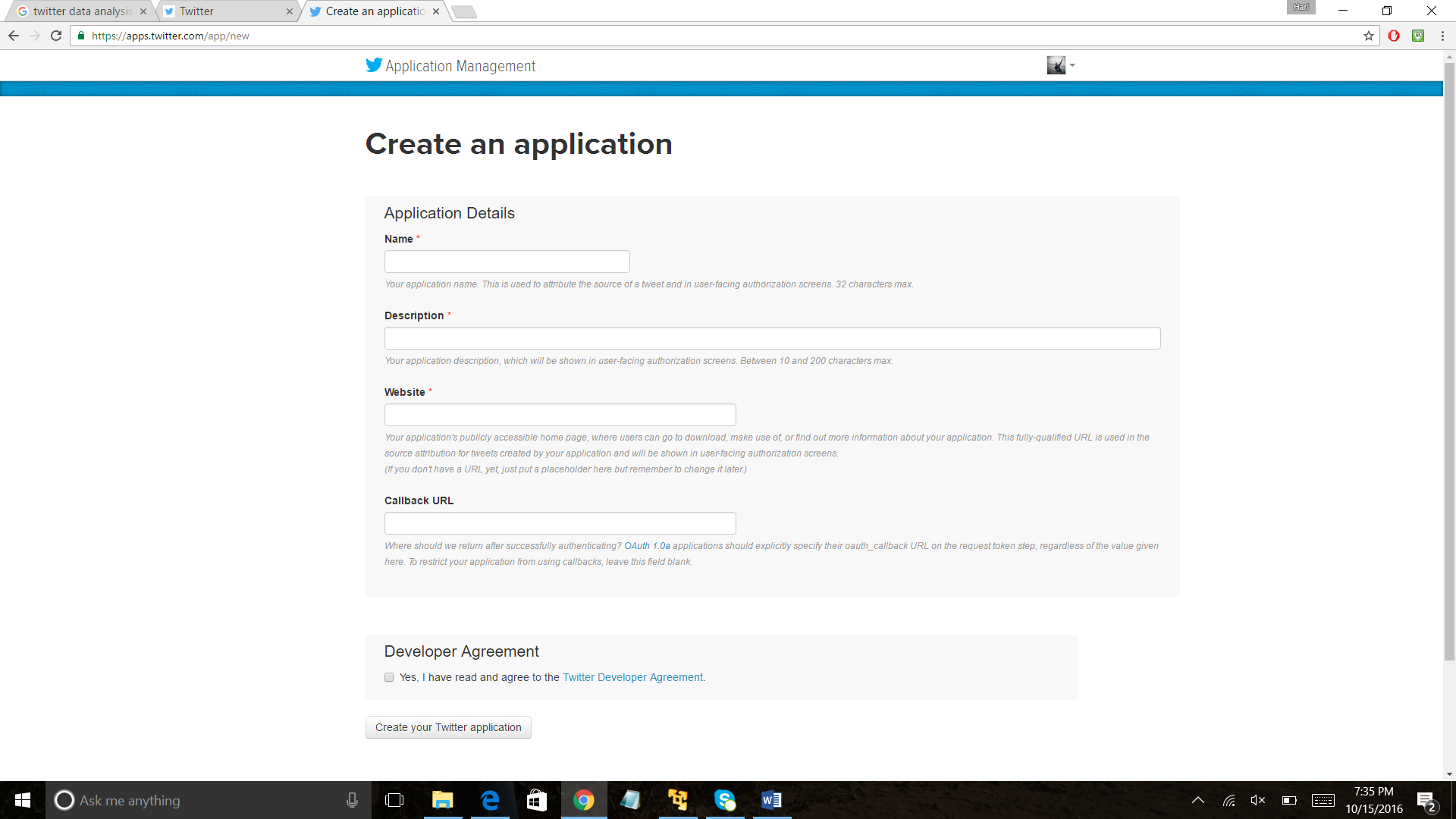
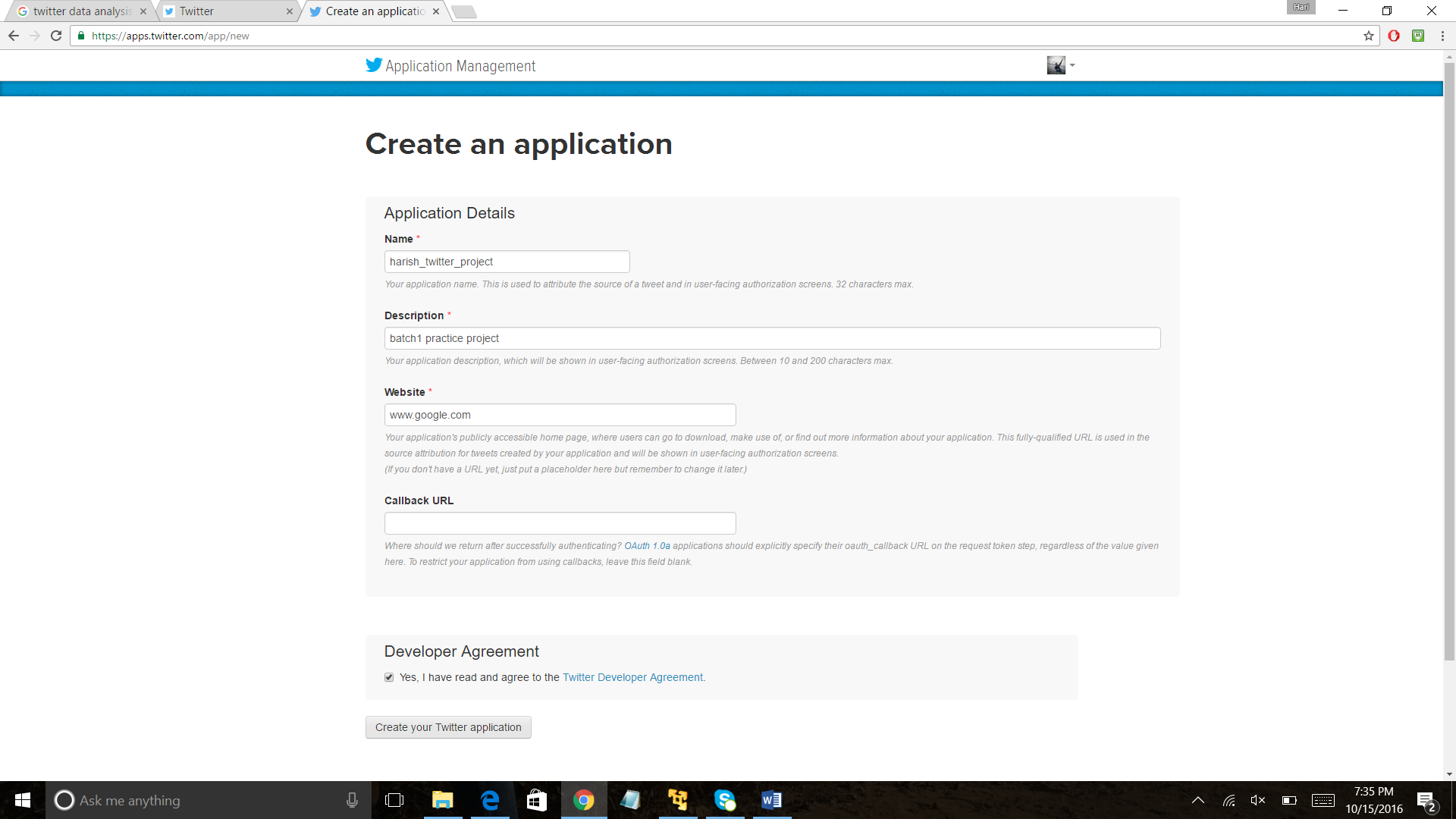
Create a twitter account and login.

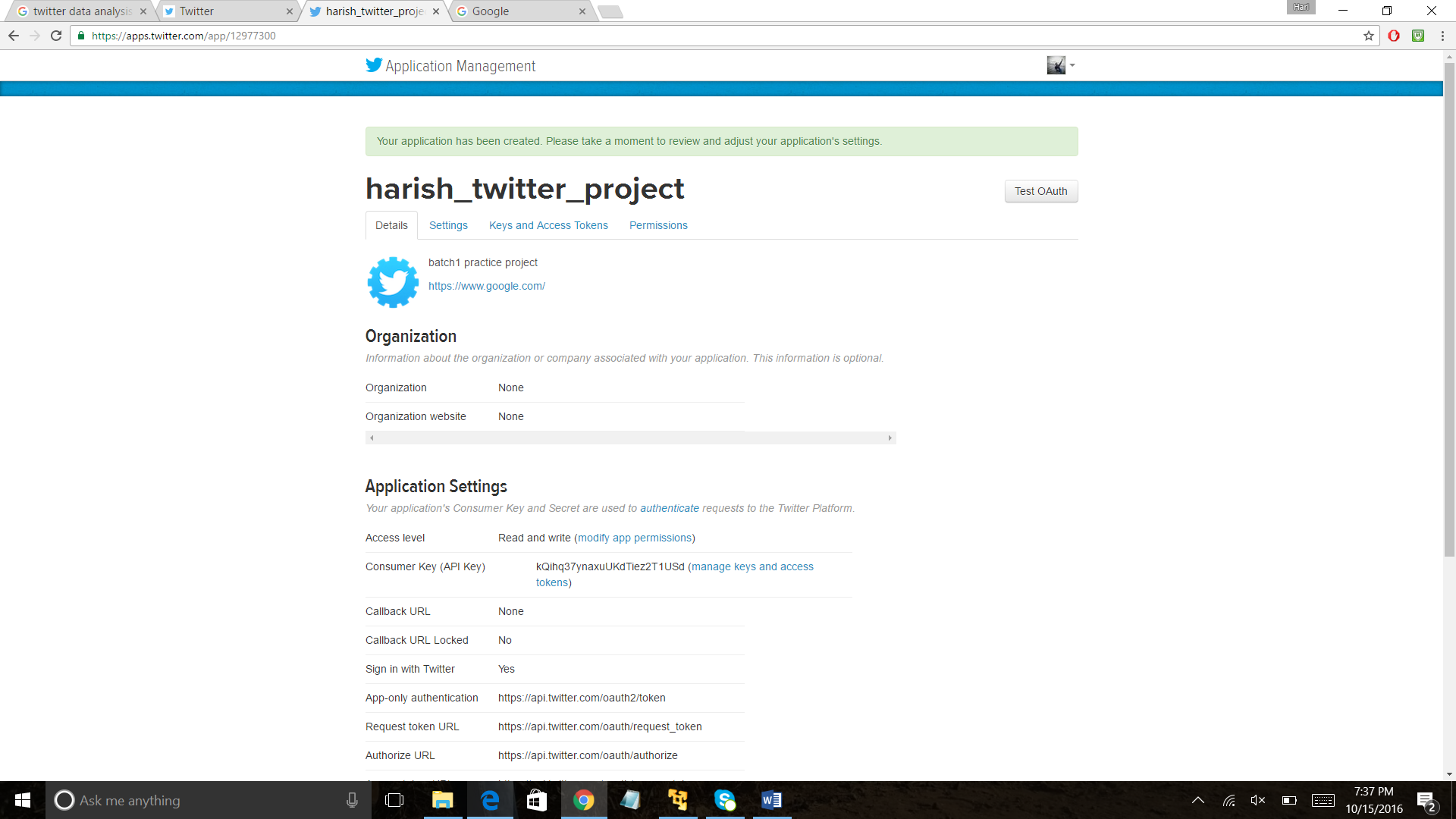


Go to apps.twitter.com and click on “Create App” as shown below in screenshots.

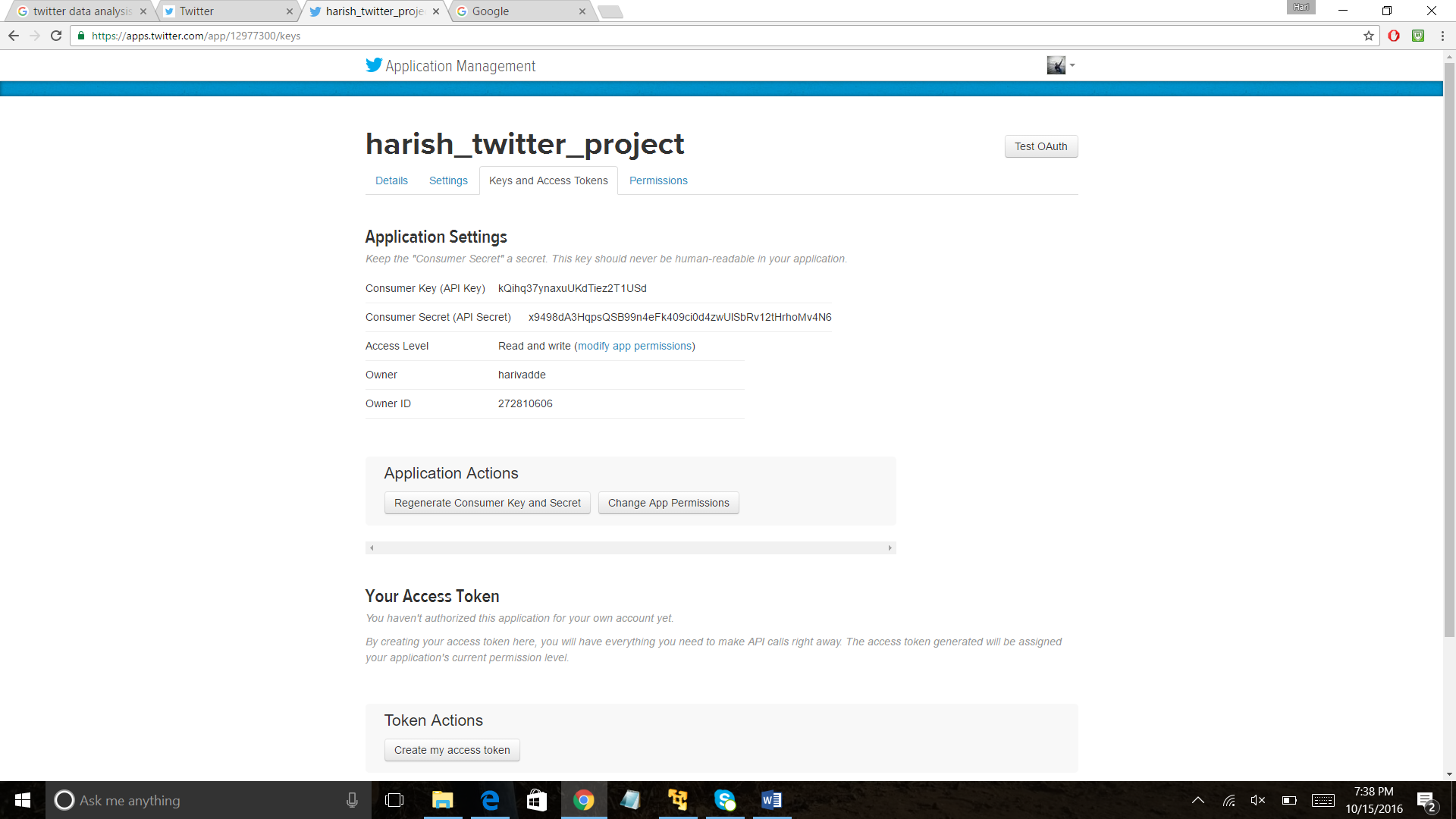




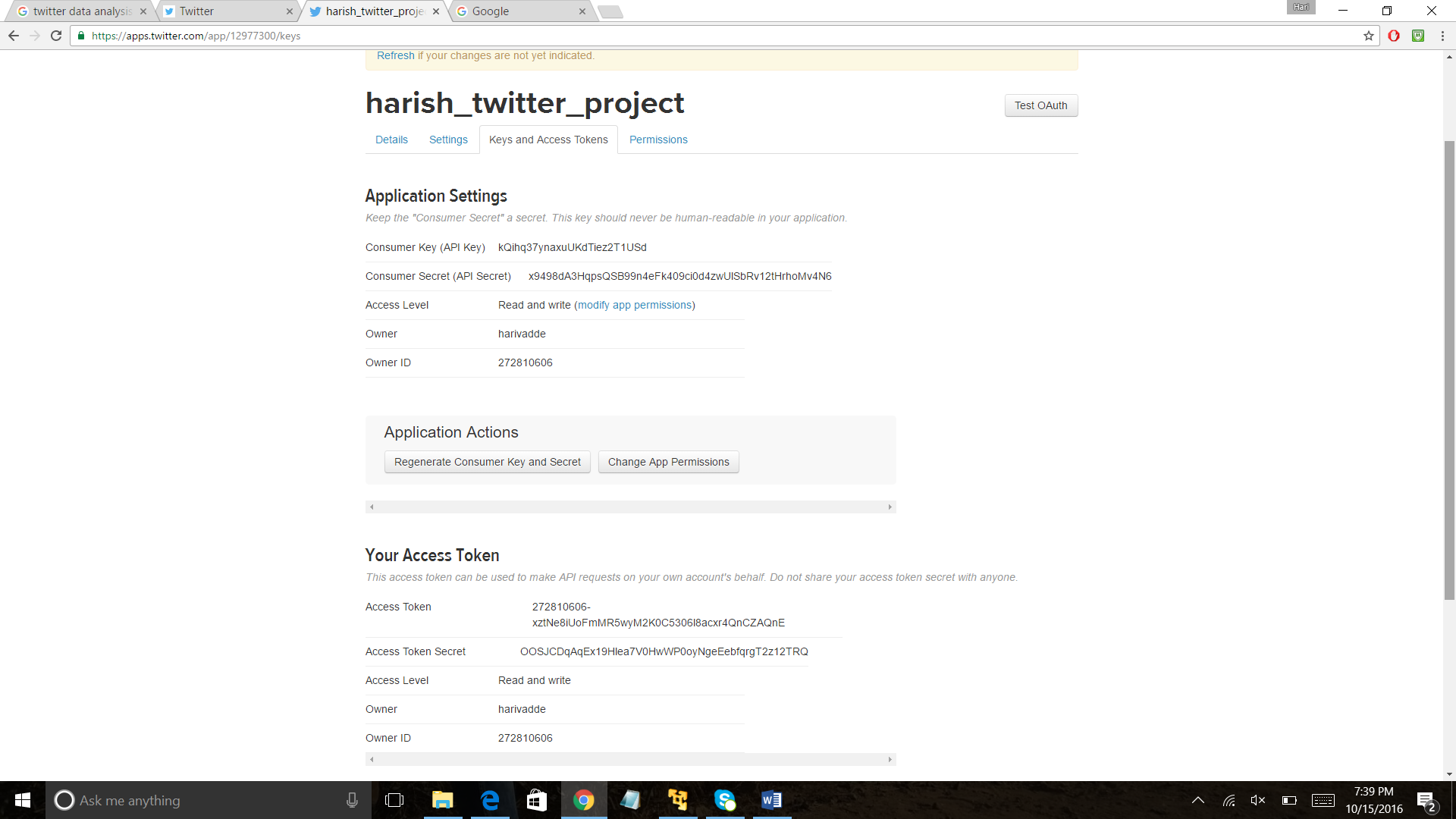




Once you have created an application, click on “keys and access tokens”



Click on create access token in the Your Access Token.



**Setting up flume agent:**

Download latest jar file called flume-sources-1.0-SNAPSHOT.jar from the below link.

<http://blog.qburst.com/2013/10/capturing-tweets-using-apache-flume/>

Copy and Paste the jar file in the flume lib folder.

sudo cp flume-sources-1.0-SNAPSHOT.jar /usr/local/apache-flume-1.6.0-bin/lib

sudo chmod +r /usr/local/apache-flume-1.6.0-bin/lib/ flume-sources-1.0-SNAPSHOT.jar

Update flume-env.sh.

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-i386

FLUME\_CLASSPATH="/usr/local/apache-flume-1.6.0-bin/lib/flume-sources-1.0-SNAPSHOT.jar"

**Create a HDFS directory to get twitter data:**

**$ hadoop fs -mkdir /TwitterProject**

Edit flume conf file “flume.conf” to change twitter keys and tokens and hdfs path. Use twitter access tokens from twitter account. Change HDFS path to “/TwitterProject/%Y/%m/%d/%H/”

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = kQihq37ynaxuUKdTiez2T1USd

TwitterAgent.sources.Twitter.consumerSecret = x9498dA3HqpsQSB99n4eFk409ci0d4zwUlSbRv12tHrhoMv4N6

TwitterAgent.sources.Twitter.accessToken = 272810606-xztNe8iUoFmMR5wyM2K0C5306l8acxr4QnCZAQnE

TwitterAgent.sources.Twitter.accessTokenSecret = OOSJCDqAqEx19Hlea7V0HwWP0oyNgeEebfqrgT2z12TRQ

TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientiest, business intelligence, mapreduce, data warehouse, data warehousing, m$

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:9000/TwitterProject/%Y/%m/%d/%H/

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

**Goto Flume conf file and execute following command:**

**$flume-ng agent -n TwitterAgent --conf ./conf/ -f flume.conf**

**Hive Setup:**

add jar using add jar command as shown below

Hive> add jar hive-serdes-1.0-SNAPSHOT.jar;

Need to create External hive table.

**Hive External Table:**

create external table if not exists twitter\_retweet\_analysis11(

filter\_level STRING,

retweeted BOOLEAN,

possibly\_sensitive BOOLEAN,

truncated BOOLEAN,

lang STRING,

in\_reply\_to\_status\_id STRING,

created\_at STRING,

favorite\_count INT,

place STRUCT<id: STRING,

bounding\_box:STRUCT<type: STRING,coordinates:ARRAY<STRUCT<items:INT>>>,

place\_type: STRING,

name: STRING,

attributes:STRING,

country\_code: STRING,

url: STRING,

country: STRING,

full\_name: STRING

>,

coordinates STRING,

text STRING,

contributors STRING,

geo STRING,

quoted\_status\_id\_str STRING,

entities STRUCT<

symbols: ARRAY<STRING>,

urls: ARRAY<STRUCT<expanded\_url: STRING, indices:ARRAY<STRUCT<items: INT>>, display\_url: STRING, url: STRING>>,

hashtags: ARRAY<STRUCT<text: STRING, indices:ARRAY<STRUCT<items: INT>>>>,

user\_mentions: ARRAY<STRING>

>,

quoted\_status\_id INT,

is\_quote\_status BOOLEAN,

source STRING,

favorited BOOLEAN,

in\_reply\_to\_user\_id INT,

retweet\_count INT,

id\_str STRING,

`user` STRUCT< location: STRING,

default\_profile: BOOLEAN,

profile\_background\_tile: BOOLEAN,

statuses\_count: INT,

lang: STRING,

profile\_link\_color: STRING,

profile\_banner\_url: STRING,

id: INT,

`following`: STRING,

protected: BOOLEAN,

favourites\_count: INT,

profile\_text\_color: STRING,

verified: BOOLEAN,

description: STRING,

contributors\_enabled: BOOLEAN,

profile\_sidebar\_border\_color: STRING,

name: STRING,

profile\_background\_color: STRING,

created\_at: STRING,

default\_profile\_image: BOOLEAN,

followers\_count: INT,

profile\_image\_url\_https: STRING,

geo\_enabled: BOOLEAN,

profile\_background\_image\_url: STRING,

profile\_background\_image\_url\_https: STRING,

follow\_request\_sent: STRING,

url: STRING,

utc\_offset: INT,

time\_zone: STRING,

notifications: STRING,

profile\_use\_background\_image: BOOLEAN,

friends\_count: INT,

profile\_sidebar\_fill\_color: STRING,

screen\_name: STRING,

id\_str: STRING,

profile\_image\_url: STRING,

listed\_count: INT,

is\_translator: BOOLEAN>

)

ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'

location '/twitter\_analysis1/';

**Loading Data into Hive Table:**

load data inpath '/TwitterProject/2016/10/15/21' overwrite into table twitter\_retweet\_analysis1;

**Analysis**: Which Twitter users get the most retweets? Who is influential within our industry?

**hive> select t.retweeted\_screen\_name,sum(retweets) AS total\_retweets,count(\*) AS tweet\_count**

**> from (select `user`.screen\_name as retweeted\_screen\_name,text,max(retweet\_count) as retweets FROM twitter\_retweet\_analysis GROUP BY `user`.screen\_name,text)**

**t group by t.retweeted\_screen\_name order by total\_retweets DESC, tweet\_count ASC limit 10;**